

AD-A265 098



ARO Report 93-1

Handwritten mark resembling a stylized 'D' or 'B'.

TRANSACTIONS OF THE TENTH ARMY

CONFERENCE ON APPLIED MATHEMATICS

AND COMPUTING



DTIC
ELECTE
MAY 28 1993
S A D

Approved for public release; distribution unlimited.
The findings in this report are not to be construed as
an official Department of the Army position, unless
so designated by other authorized documents.

93-12045



93 5 27 048 Sponsored by

The Army Mathematics Steering Committee

on behalf of

**THE ASSISTANT SECRETARY OF THE ARMY FOR
RESEARCH, DEVELOPMENT, AND ACQUISITION**

U.S. ARMY RESEARCH OFFICE

Report No. 93-1

March 1993

TRANSACTIONS OF THE TENTH ARMY CONFERENCE
ON APPLIED MATHEMATICS AND COMPUTING

Sponsored by the Army Mathematics Steering Committee

HOST

U.S. Military Academy
West Point, New York
16-19 June 1992

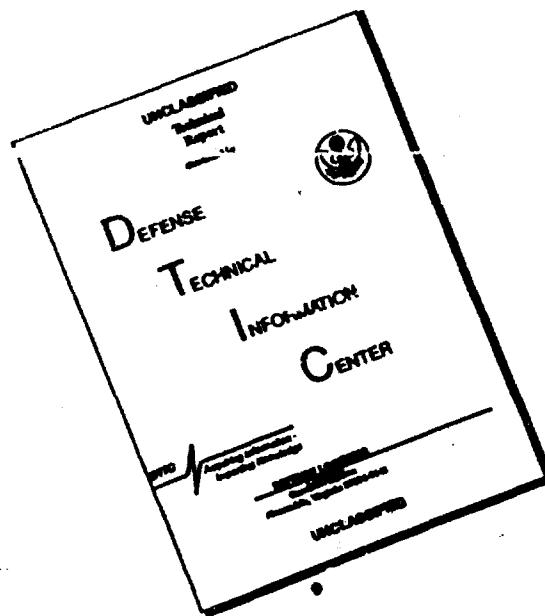
CONFIDENTIAL

Approved for public release; distributions unlimited.
The findings in this report are not to be constructed as
an official Department of the Army position, unless so
designated by other authorized documents.

U.S. Army Research Office
P. O. Box 12211
Research Triangle Park, NC 27709-2211

Accession For	
NTIS	CRA&I <input checked="" type="checkbox"/>
DTIC	TAB <input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By	
Distribution /	
Availability Codes	
Dist	Availability or Special
A-1	

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

FOREWORD

The Tenth Army Conference on Applied Mathematics and Computing was held at the U.S. Military Academy, West Point, on 16-19 June 1992. This is the third time the Military Academy has served as the host for this series of Army conferences. For each of these meetings the heads of the Department of Mathematics served as Chairpersons on Local Arrangements. Colonel Frank Giordano served twice in this capacity. For the tenth conference, he was assisted in this task by Lieutenant Colonel Scott Crawford and Captain Rick Stevens. These individuals are to be commended for their efforts in coordinating all the details required to conduct this large successful scientific meeting.

The 1992 conference was attended by more than 80 scientists and engineers representing various Army agencies and academia. The meeting featured seven invited speakers. These general talks covered several topics of current interest, including natural language processing, wavelet analysis, variational methods, small sample asymptotics, computational fluid dynamics, digital control programs, and parallel programming. The names of these speakers, together with the titles of their addresses are listed below. The second part of the program consisted of special sessions on topics such as computational algebraic geometry, mathematical aspects of material sciences, scalability in high performance computing, and robust control and nonlinear systems. In addition, about 50 contributed papers were presented by both Army and Academic participants.

SPEAKER AND AFFILIATION

TITLE OF ADDRESS

Professor Anil Nerode and
Alexander Yakhnis
Cornell University

Extraction of Digital Control
Programs from Specifications
for Continuous Plants

Professor Ken Kennedy
Rice University

Architecture-Independent
Parallel Programming Support
in Fortran D

Professor C. R. Rao
Pennsylvania State

Current Trends of Research in
Statistics: Robustness, Small
Sample Asymptotics & Resampling

Professor Aravind K. Joshi
University of Pennsylvania

Natural Language Professing

Professor Charles K. Chui
Texas A&M University

Wavelet Analysis and Its
Applications

Professor David Kinderlehrer
Carnegie Mellon University

Variational Methods for the
Materials Sciences

SPEAKER AND AFFILIATIONTITLE OF ADDRESS

Professor Paul Woodward
University of Minnesota

Visualization in Computational
Dynamics

The evening events provided by the host installation added a great deal to this conference: On Tuesday evening there was a bus tour of West Point, on the following evening the banquet was held, and the Thursday evening workshop, put on by the members of the Mathematics Division, introduced some attendees of the conference to computer algebra.

This conference is part of a continuing program of Army-wide symposia held under the auspices of the Army Mathematics Steering Committee (AMSC) to promote better communication between Army scientists and the Army Research Office investigators. In order that this mission be accomplished, a large number of scientists had to expend a great deal of effort. The members of the AMSC would like to thank all these individuals for their excellent presentations and their valuable contributions to the field of science.

TABLE OF CONTENTS*

<u>Title</u>	<u>Page</u>
Foreword	iii
Table of Contents	v
Agenda	ix
Toward a New Method of Decoding Algebraic Codes Using Grobner Bases A. Brinton Cooper, III	1
An Algorithm for the Computation of Grobner Bases W.W. Adams, A. Boyle and P. Loustau	13
Long Time Behavior of a Numerical Approximation to a Nonlinear Evolution Problem in Viscelasticity Donald A. French	23
Condensation of Three-Dimensional Finite Elements To Solve Problems of Wave Propagation David W. Sykora and Jose M. Roeset	29
Generalized Stroh Formalism for Anisotropic Elasticity for General Boundary Conditions T. C. T. Ting and M. Z. Wang	53
Computation of Microstructure Utilizing Young Measure Representations R. A. Nicolaides and Noel J. Walkington	57
Kinetically Driven Elastic Phase Boundary Motion Activated by Concurrent Dynamic Pulses Jiehliang Lin and Thomas J. Pence	69
A Function Whose Values are Integers, II J. Arkin, D. C. Arney and E. H. Luchins	87

*This Table of Contents lists only the papers that are published in this Technical Manual. For a list of all the papers presented at the tenth Army Conference on Applied Mathematics and Computing, see the Agenda.

<u>Title</u>	<u>Page</u>
Reverse Digit Constructions of Perfect, Magic, and Doubly Magic Cubes J. Arkin, D.C. Arney, F. R. Giordano and R. Kolb	91
Analytic Roots of the Period Three Quadratic Recursion Polynomial Harry J. Auvermann	103
A Godunov Scheme for Elasto-Plasticity J. W. Grove, B.J. Plohr, D.H. Sharp, and F. Wang	113
The Korteweg Theory of Cappilarity and the Phase Transition Problems Harumi Hattori	131
Singular Value Computation on a Fat-Tree Network Tong J. Lee, Franklin T. Luk and Daniel L. Boley	143
A Maximal Invariant Framework for Adaptive Detection Sandip Bose and Allan Steinhardt	153
Scalable Software Tools for Adaptive Science Computation B. K. Szymanski, C. Ozturan and J. E. Flaherty	159
Automated Interpretation of Topographic Maps T. Cronin	173
Current Trends of Research in Statistics Small Sample Asymptotics Resampling Techniques and Robustness C. Radharkrishna Rao	195
Variational Theory of Motion of Curved, Twisted and Extensible Elastic Rods Iradj Tadjbakhsh and Dimitris C. Lagoudas	221
Viscohyperelasticity A. R. Johnson, C. J. Quigley and C. E. Freese	235
Multigrid Algorithms for the First Biharmonic Problem: Robustnes M. R. Hanisch	257

<u>Title</u>	<u>Page</u>
Simulated Annealing Algorithms for Continuous Optimization S. B. Gelfand, P. C. Doerschuk and M. Nahhas-Mohandes	273
Response of Cylindrical Section to an Explosive Blast Aaron Das Gupta	283
Development of a Computational Method for Conventional Weapons Analysis of Buried Structures James T. Baylot	291
Adaptive Grids for the Hull Hydrodynamics Code C. Wayne Mastin	313
Finite Dimensional Estimation Algebras of Maximal Rank With Dimension of State Space Equal to 3 Stephen S. T. Yau, Jie Chen and Chi-Wah Leung	337
Hybrid Optimal Control of Turret-Gun System J. L. Zhang, L. S. Shieh and N. P. Coleman	345
Wavelet Analysis and Its Applications Charles K. Chui	359
Design and Analysis of Scalable Parallel Algorithms Vipin Kumar	373
Dynamical Systems in Asymmetric Space and Time Richard A. Weiss	385
Slow and Ultrafast Wave Propagation Processes Richard A. Weiss	417
Clean Fission Nuclear Reactors Richard A. Weiss	463
Thermonuclear Reactions in Strong Gravitational Fields Richard A. Weiss	555
Some Methods of Analysis in the Study of Microstructure David Kinderhrer	613

<u>Title</u>	<u>Page</u>
Classical Finite Element Method for Transient Three Demensional Heat Conduction Rao Yalamanchili and S. Yalamanchili	635
Non-Collocated Motion Control of A Flexible Beam Based On A Delayed Adaptive Inverse Method D. Wang, G. Yang, M. Donath, M. Mattice and N. Coleman	643
Preliminary Mu-synthesis Design for the ATB-1000 D. Evans, D. Bugajski and A. Tannenbaum	671
An Introduction to the Trajectory Pattern Method and Its Control Application F. Tangerman and J. Rastegar	683
Visualization of Dynamic Soil-Structure Interaction Analysis J. Baca, R. L. Hall and D. H. Nelson	699
Advanced Computer Modeling of Meterological Effects upon Artillery Projectile Flight A. J. Blanco and S. J. Edwards	707
Experimental Techniques for Scientific Data Interpretation Charles S. Jones and Julia A. Baca	717
Flow Computation about Army Projectiles and Missiles Using Multi-Zone Grids on Parallel Computer Architectures Dr. N. Patel, J. Clarke and M. Coleman	729
Application of Finite Element, Grid Generaltion, and Scientific Visualizations Techniques to 2-D and 3-D Seepage and Groundwater modeling Fred T. Tracy and Camille A. Issa	743
List of Conference Attendees	761

TENTH ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

Host

U.S. Military Academy, West Point, New York

16-19 June 1992

AGENDA

Tuesday, 16 June 1992

0745 - 1600 Registration - Thayer Hall, Room 342

0815 - 0830 Opening Remarks - Thayer Hall, Room 342

0830 - 0930 General Session I - Thayer Hall, Room 342

**Chairperson: Benjamin E. Cummings, U.S. Army Human
Engineering Laboratory, Aberdeen Proving Ground,
Maryland**

***EXTRACTION OF DIGITAL CONTROL PROGRAMS FROM
SPECIFICATIONS FOR CONTINUOUS PLANTS***

**Anil Nerode and Alexander Yakhnis, Cornell University,
Ithaca, New York**

0930 - 1000 Break

**1000 - 1200 Special Session 1 - Computational Algebraic Geometry
- Thayer Hall, Room 342**

**Chairperson: Kenneth D. Clark, U.S. Army Research Office,
Research Triangle Park, North Carolina**

***COMPUTATIONAL ALGEBRAIC GEOMETRY: AN EQUATION SOLVING
PERSPECTIVE***

Michael Stillman, Cornell University, Ithaca, New York

***TOWARD A NEW METHOD OF DECODING ALGEBRAIC CODES
USING GRÖBNER BASES***

**Brinton Cooper, U.S. Army Ballistic Research Laboratory,
Aberdeen Proving Ground, Maryland**

Tuesday (continued)

TRANSITIVITY IN THE THEORY OF GRÖBNER BASES

Phillipe Loustau, George Mason University, Fairfax, Virginia;
Ann Boyle, National Science Foundation, Washington, D.C.

IMPLICIT ALGEBRAIC SPLINES WITH APPLICATIONS

Chanderjit Bajaj, Purdue University, West Lafayette, Indiana

XXXXXXXXXXXXXXXXXXXXXXX

1000 - 1200

**Technical Session 1 - Approximation and Finite Element Methods
- Thayer Hall, Room 348**

**Chairperson: Royce Soanes, U.S. Army Armament, Research,
Development and Engineering Center, Watervliet,
New York**

***THE APPROXIMATION ORDER OF FINITELY GENERATED SHIFT-
INVARIANT SPACES IN $L_2(R^d)$***

Carl deBoor, University of Wisconsin, Madison, Wisconsin

***ADAPTATION OF MULTI-REGION CUBIC SPLINE FUNCTION FOR
USE WITH MARQUARDT METHOD***

Douglas R. Sommerville, Chemical Research, Development and
Engineering Center, Aberdeen Proving Ground, Maryland

***CLASSICAL FINITE ELEMENT METHOD FOR TRANSIENT THREE
DIMENSIONAL HEAT CONDUCTION***

Rao Yalamanchili, U.S. Army Armament, Research and
Development Center, Picatinny Arsenal, New Jersey, and Surya
Yalamanchili, Rutgers University, New Brunswick, New Jersey

***CONTINUOUS TIME GALERKIN METHODS FOR NONLINEAR
EVOLUTION PROBLEMS***

Donald A. French, University of Cincinnati, Cincinnati, Ohio

***CONDENSATION OF THREE-DIMENSIONAL FINITE ELEMENTS TO
SOLVE PROBLEMS OF WAVE PROPAGATION***

David W. Sykora, U.S. Army Waterways Experiment Station,
Vicksburg, Mississippi and J.M. Roesset, University of Texas
at Austin, Austin, Texas

Tuesday (continued)

1200 - 1330

Lunch

1330 - 1530

**Special Session 2A - Mathematical Aspects of Material Sciences
- Thayer Hall, Room 342**

**Chairperson: Julian Wu, U.S. Army Research Office, Research
Triangle Park, North Carolina**

***GENERALIZED STROH FORMALISM FOR ANISOTROPIC
ELASTICITY FOR GENERAL BOUNDARY CONDITIONS***

**Thomas C. T. Ting, University of Illinois at Chicago, Chicago,
Illinois**

PIEZOELECTRICITY IN COMPOSITE MATERIALS

**Marco Avellaneda, Courant Institute for Mathematical Sciences,
New York, New York**

***NON-CONVEX OPTIMIZATION AND COMPUTATION OF
MICROSTRUCTURE***

**Noel J. Walkington, Carnegie Mellon University, Pittsburgh,
Pennsylvania**

***DYNAMIC DECAY TO ENERGY MINIMUM EQUILIBRIUM STATES IN
ELASTIC MATERIALS SUSTAINING MULTIPLE PHASES***

**Thomas J. Pence and Jack Lin, Michigan State University,
East Lansing, Michigan**

XXXXXXXXXXXXXXXXXXXXX

1330 - 1530

**Technical Session 2 - Algebraic and Symbolic Methods
- Thayer Hall, Room 348**

**Chairperson: Rao Yalamanchili, U.S. Army Armament, Research
and Development Center, Dover, New Jersey**

WINDING NUMBERS AND STURM SEQUENCES

Moss Sweedler, Cornell University, Ithaca, New York

Tuesday (continued)

HOMOTOPY METHODS FOR SPARSE POLYNOMIAL SYSTEMS

Birkett Huber, Cornell University, Ithaca, New York

UNIQUE FACTORIZATIONS (base 2)

Joseph Arkin, U.S. Military Academy, West Point, New York

RESEARCHES ON THE K th POWER OF SERIES

Joseph Arkin and Edith H. Luskin, U.S. Military Academy,
West Point, New York

STRAIGHTENING EUCLIDEAN INVARIANTS

John P. Dalbec, Cornell University, Ithaca, New York

***ANALYTIC ROOTS OF THE PERIOD THREE QUADRATIC
RECURSION POLYNOMIAL***

Harry J. Auvermann, U.S. Army Atmospheric Sciences
Laboratory, White Sands Missile Range, New Mexico

1530 - 1600

Break

1600 - 1700

General Session II - Thayer Hall, Room 342

**Chairperson: Kenneth D. Clark, U.S. Army Research Office,
Research Triangle Park, North Carolina**

***ARCHITECTURE-INDEPENDENT PARALLEL PROGRAMMING
SUPPORT IN FORTRAN D***

Ken Kennedy, Rice University, Houston, Texas

XXXXXXXXXXXXXXXXXXXXX

Wednesday, 17 June 1992

0800 - 1600

Registration - Thayer Hall, Room 342

0830 - 1030

**Special Session 2B - Mathematical Aspects of Material Sciences
- Thayer Hall, Room 342**

**Chairperson: Julian Wu, U.S. Army Research Office,
Research Triangle Park, North Carolina**

Wednesday (continued)

TOTAL ABSORPTION IN ELASTIC MEDIA

William W. Hager and Dongxing Wang, University of Florida,
Gainesville, Florida and, Rouben Rostamian, University of
Maryland, Baltimore County, Catonsville, Maryland

**A CONSERVATIVE EULERIAN SCHEME FOR MODELING FINITE
DEFORMATION IN ELASTIC-PLASTIC SOLIDS**

Bradley J. Plohr and Feng Wang, SUNY at Stony Brook,
Stony Brook, New York

ON SLOW MOTIONS IN A PHASE TRANSITION PROBLEM

Harumi Hattori, West Virginia University, Morgantown, WV

**LONGITUDINAL WAVES WITH RADIAL MOTION IN A LONG ROD OF
ELASTIC-PLASTIC MATERIAL**

Peter C.T. Chen and Joseph E. Flaherty, Benet Laboratories,
Watervliet, New York

xxxxxxxxxxxxxxxxxxxxxx

0830 - 1030

**Technical Session 3 - Parallel and Adaptive Algorithms
- Thayer Hall, Room 348**

**Chairperson: J. Michael Coyle, U.S. Army Armament, Research,
Development and Engineering Center, Watervliet,
New York**

**ON SCALABLE IMPLEMENTATIONS OF A FAST HANKEL
ALGORITHM**

Tong J. Lee, Cornell University, Ithaca, New York and
Franklin T. Luk, Rensselaer Polytechnic Institute, Troy,
New York

**A MAXIMAL INVARIANT FRAMEWORK FOR ADAPTIVE DETECTION
WITH ARRAYS**

Sandip Rose and Allen O. Steinhardt, Cornell University,
Ithaca, New York

**SCALABLE SOFTWARE TOOLS FOR ADAPTIVE SCIENTIFIC
COMPUTATIONS**

Boleslaw Szymanski, Rensselaer Polytechnic Institute, Troy,
New York

Wednesday (continued)

AUTOMATED INTERPRETATION OF TOPOGRAPHIC MAPS

Terrence M. Cronin, CECOM, Center for Signals Warfare, Vint Hill Farms Station, Warrenton, Virginia

ALGORITHMS FOR PARAMETER IDENTIFICATION IN ODE's

John E. Dennis, Jr., Guangye Li, and Karen A. Williamson, Rice University, Houston, Texas

1030 - 1100

Break

1100 - 1200

General Session III - Thayer Hall, Room 342

Chairperson: Gerald R. Andersen, U.S. Army Research Office, Research Triangle Park, North Carolina

CURRENT TRENDS OF RESEARCH IN STATISTICS: ROBUSTNESS, SMALL SAMPLE ASYMPTOTICS AND RESAMPLING

C.R. Rao, Pennsylvania State University, University Park, Pennsylvania

1200 - 1330

Lunch

1330 - 1530

**Special Session 2C - Mathematical Aspects of Material Sciences
-Thayer Hall, Room 342**

Chairperson: Peter C.T. Chen, U.S. Army Armament, Research, Development and Engineering Center, Watervliet, New York

FINITE ELEMENT MODELING OF SOME PROBLEMS

J.R. Whiteman, S. Shaw and M.K. Warby, Brunel University, Uxbridge, Middlesex, United Kingdom

VARIATIONAL THEORY OF DYNAMIC DEFORMATIONS OF THE CURVED AND TWISTED ELASTICA

Iradj Tadjbakhsh and Dimitris C. Lagoudas, Rensselaer Polytechnic Institute, Troy, New York

VISCOHYPERELASTICITY

Arthur Johnson and C.J. Quigley, Materials Technology Laboratory, Watertown, Massachusetts

Wednesday (continued)

***SHOCK DAMAGE TO SENSITIVE COMPONENTS IN ARMORED
VEHICLES***

J.M. Santiago, A. Das Gupta, and H.L. Wisniewski,
Ballistic Research Laboratory, Aberdeen Proving Ground,
Maryland

XXXXXXXXXXXXXXXXXXXXXX

1330 - 1530

**Technical Session 4 - Dynamical Systems and Differential
Equations I - Thayer Hall, Room 348**

**Chairperson: John Vasilakis, U.S. Army Armament, Research,
Development and Engineering Center, Watervliet,
New York**

IRREGULAR SHOCK REFRACTIONS AT FLUID INTERFACES

John W. Grove and L.F. Henderson, SUNY at Stony Brook,
Stony Brook, New York

***MULTIGRID ALGORITHMS FOR THE FIRST BIHARMONIC
PROBLEM***

Mark Hanisch, Cornell University, Ithaca, New York

***A NON-REFLECTING BOUNDARY CONDITION FOR A ONE-WAY
EQUATION***

Major J.S. Robertson, U.S. Military Academy, West Point,
New York

***SOLUTION OF DIFFRACTION PROBLEMS VIA VARIATION OF THE
BOUNDARY***

Fernando Reitich, Carnegie Mellon University, Pittsburgh,
Pennsylvania

***ON THE COMPLEXITY OF CERTAIN SOLUTIONS TO THE
HELMOHOLTZ EQUATION***

Major J.S. Robertson, U.S. Military Academy, West Point,
New York

Wednesday (continued)

1330 - 1530 Poster Session - Thayer Hall, Room 343

***AUTOMATED INTERPRETATION OF TOPOGRAPHIC MAPS USING AN
APPLE MACINTOSH COMPUTER***

Terence M. Cronin, CECOM, Center for Signals Warfare, Vint
Hill Farms Station, Warrenton, Virginia

***STATISTICAL INFERENCE ABOUT DISCRETE PULSES OBSCURED
BY CONVOLUTION AND/OR RANDOM OBSERVATIONAL ERROR***

Arthur P. Dempster, Xianghui Chen, Jun Liu, and Patricia M.
Meehan, Harvard University, Cambridge, Massachusetts

***SIMULATED ANNEALING ALGORITHMS FOR CONTINUOUS
OPTIMIZATION***

Saul B. Gelfand, Purdue University, West Lafayette, Indiana

1530 - 1600 Break

1600 - 1700 General Session IV - Thayer Hall, Room 342

**Chairperson: David W. Hislop, U.S. Army Research Office,
Research Triangle Park, North Carolina**

NATURAL LANGUAGE PROCESSING

Aravind K. Joshi, University of Pennsylvania, Philadelphia,
Pennsylvania

xxxxxxxxxxxxxxxxxxxxxxxx

Thursday, 18 June 1992

0800 - 1600 Registration - Thayer Hall, Room 342

**0830 - 1030 Technical Session 5 - Applied Finite Element Computations
- Thayer Hall, Room 342**

**Chairperson: C.J. Quigley, U.S. Army Materials Technology Laboratory,
Watertown, Massachusetts**

***STRUCTURAL ANALYSIS OF A CYLINDRICAL GUN TUBE SECTION
SUBJECTED TO AN INTERNAL EXPLOSIVE BLAST***

Aaron Das Gupta, U.S. Army Ballistic Research Laboratory,
Aberdeen Proving Ground, Maryland

Thursday (continued)

***DEVELOPMENT OF A COMPUTATIONAL METHOD FOR CONVENTIONAL
WEAPONS ANALYSIS OF BURIED STRUCTURES***

James T. Baylot, U.S. Army Engineer Waterways Experiment Station,
Vicksburg, Mississippi

***HIGH SPEED FLOW SIMULATION USING A HIGHER-ORDER LOCAL
PROJECTION FINITE ELEMENT METHOD ON MASSIVELY PARALLEL
COMPUTERS***

Bernardo Cockburn, University of Minnesota, Minneapolis, Minnesota,
and Nisheeth Patel and Jerry Clarke, U.S. Army Ballistic Research
Laboratory, Aberdeen Proving Ground, Maryland

ADAPTIVE GRIDS OF HULL HYDRODYNAMICS CODE

C. Wayne Mastin, U.S. Army Engineer Waterways Experiment Station,
Vicksburg, Mississippi

xxxxxxxxxxxxxxxxxxxxxxxx

**0830 - 1030 Technical Session 6 - Nonlinear Systems and Adaptive Control
- Thayer Hall, Room 348**

**Chairperson: William Jackson, U.S. Army Tank-Automotive Command,
Warren, Michigan**

***SMOOTH TIME-PERIODIC FEEDBACK SOLUTIONS FOR NONHOLONOMIC
MOTION PLANNING***

L. Gurvits and Z.X. Li, Courant Institute of Mathematical Sciences,
New York University, New York, New York

***CLASSIFICATION OF FINITE DIMENSIONAL ESTIMATION ALGEBRAS:
LOW DIMENSIONAL CASE***

Stephen Yau, Chi-Wah Leung, University of Illinois at Chicago,
Chicago, Illinois

***ADAPTIVE GUNNER MODELING FOR APPLICATION TO ARMORED
VEHICLE TECHNOLOGY***

John Groff and Peter Pazio, U.S. Army Ballistic Research Laboratory,
Aberdeen Proving Ground, Maryland

Thursday (continued)

MODELING AND CONTROL ISSUES OF A TWIN LIFT HELICOPTER SYSTEM

J.V.R. Prasad, Georgia Institute of Technology, Atlanta, Georgia

HYBRID OPTIMAL CONTROL OF TURRET-GUN SYSTEM

L.S. Shieh and J.L. Zhang, University of Houston, Houston, Texas,
and Norman Coleman, U.S. Army Armament, Research and
Development Center, Picatinny Arsenal, New Jersey

1030 - 1100 Break

1100 - 1200 General Session V - Thayer Hall, Room 342

**Chairperson: Frank A. Giordano, U.S. Military Academy, West Point,
New York**

WAVELET ANALYSIS AND ITS APPLICATIONS

Charles K. Chui, Texas A&M University, College Station, Texas

1200 - 1330 Lunch

**1330 - 1530 Special Session 3 - Scalability in High Performance Computing
- Thayer Hall, Room 342**

**Chairperson: Kenneth D. Clark, U.S. Army Research Office,
Research Triangle Park, North Carolina**

SCALABILITY ISSUES IN INTERCONNECTION NETWORKS

Frank T. Leighton, Massachusetts Institute of Technology, Cambridge,
Massachusetts

SCALABLE PARALLELISM FOR PDE COMPUTATIONS

John R. Rice, Purdue University, West Lafayette, Indiana

***ANALYZING SCALABILITY OF PARALLEL ALGORITHMS AND
ARCHITECTURES***

Vipin Kumar, University of Minnesota, Minneapolis, Minnesota

XXXXXXXXXXXXXXXXXXXXX

Thursday (continued)

1330 - 1530 Technical Session 7 - Dynamical Systems and Differential Equations II - Thayer Hall, Room 348

Chairperson: J.S. Robertson, U.S. Military Academy, West Point, New York

FRONT-PROPAGATION IN REACTION-DIFFUSION EQUATIONS

Halil Mete Soner, Carnegie Mellon University, Pittsburgh, Pennsylvania

NUMERICAL TREATMENT OF RANDOM DIFFERENTIAL EQUATIONS

G.S. Ladde, University of Texas at Arlington, Arlington, Texas, and
S. Sathananthan, Jarvis Christian College, Hawkins, Texas

DIAGONALIZATION AND STABILITY OF TWO TIME-SCALE SINGULARLY PERTURBED LINEAR INTEGRO-DIFFERENTIAL EQUATIONS

G.S. Ladde, University of Texas at Arlington, Arlington, Texas, and
S. Sathananthan, Jarvis Christian College, Hawkins, Texas

1) DYNAMICAL SYSTEMS IN ASYMMETRIC SPACE AND TIME

2) SLOW AND ULTRAFAST WAVE PROPAGATION

Richard A. Weiss, U.S. Army Engineer Waterways Experiment Station,
Vicksburg, Mississippi

CONSIDERATION OF PHASE TRANSFORMATIONS OF SHEAR BANDS IN A DYNAMICALLY LOADED STEEL BLOCK

Z.G. Zhu and R.C. Batra, University of Missouri-Rolla, Rolla, Missouri

1530 -1600 Break

1600 - 1700 General Session VI - Thayer Hall, Room 342

**Chairperson: Julian Wu, U.S. Army Research Office,
Research Triangle Park, North Carolina**

VARIATIONAL METHODS FOR THE MATERIALS SCIENCES

David Kinderlehrer, Carnegie Mellon University, Pittsburgh,
Pennsylvania

XXXXXXXXXXXXXXXXXXXXX

Friday, 19 June 1992

0800 - 1200 Registration - Thayer Hall, Room 342

**0830 - 1030 Special Session 4 - Robust Control and Nonlinear Systems
- Thayer Hall, Room 342**

**Chairperson: Norman Coleman, U.S. Army Armament, Research,
Development and Engineering Center, Picatinny Arsenal,
New Jersey**

***NON-COLLOCATED MOTION CONTROL OF A FLEXIBLE BEAM BASED ON
A DELAYED ADAPTIVE INVERSE METHOD***

David S. Wang, Guo-Ben Yang, and Max Donath, University of
Minnesota, Minneapolis, Minnesota, and Mike Mattice and Norman
Coleman, U.S. Army Armament, Research, Development and
Engineering Center, Picatinny Arsenal, New Jersey

CONTROL DESIGN FOR FLEXIBLE STRUCTURES

Karashad Khorrami, Polytechnic University, Brooklyn, New York

***A NONLINEAR FEEDBACK CONTROL STRATEGY BASED ON TRAJECTORY
PATTERNS***

Dale Enns, Honeywell Systems Research Center, Minneapolis,
Minnesota, and Allen Tannenbaum, University of Minnesota,
Minneapolis, Minnesota

***ADVANCED NONLINEAR CONTROL THEORY FOR GUN CONTROL/
STABILIZATION***

Jay Rastegar and Folkert Tangerman, SUNY at Stony Brook,
Stony Brook, New York

xxxxxxxxxxxxxxxxxxxxxxxx

**0830 - 1030 Technical Session 8 - Computer Modeling and Scientific Visualization
- Thayer Hall, Room 348**

**Chairperson: Herbert Cohen, U.S. Army Material Systems Analysis
Activity, Aberdeen Proving Ground, Maryland**

VISUALIZATION OF DYNAMIC SOIL-STRUCTURE INTERACTION ANALYSIS

Robert L. Hall, Julia A. Baca, and Donald Nelson, U.S. Army Engineer
Waterways Experiment Station, Vicksburg, Mississippi

Friday (continued)

***ADVANCED COMPUTER MODELING OF METEOROLOGICAL EFFECTS
UPON ARTILLERY PROJECTILE FLIGHT***

Abol J. Blanco, U.S. Army Atmospheric Sciences Laboratory,
White Sands Missile Range, New Mexico

SCIENTIFIC VISUALIZATION OF FLOW FIELDS

Charles S. Jones and Julia A. Baca, U.S. Army Engineer Waterways
Experiment Station, Vicksburg, Mississippi

***THE GENERALIZED BALANCED TENNARY (GBT) PROPOSED FOR
APPLICATION SCIENTIFIC VISUALIZATION***

Robert E. DeKinder, U.S. Army Atmospheric Sciences Laboratory,
White Sands Missile Range, New Mexico, and John R. Barns,
Computer Sciences Corporation, White Sands Missile Range Support
Office, Las Cruces, New Mexico

***FLOW COMPUTATION ABOUT ARMY PROJECTILES AND MISSILES USING
MULTI-ZONE PARALLEL COMPUTER ARCHITECTURES***

Nisheeth Patel, Jerry Clarke, and Monte Coleman, U.S. Army Ballistic
Research Laboratory, Aberdeen Proving Ground, Maryland

***APPLICATION OF FINITE ELEMENT, GRID GENERATION, AND SCIENTIFIC
VISUALIZATION TECHNIQUES***

Fred Tracy, U.S. Army Engineer Waterways Experiment Station,
Vicksburg, Mississippi

1030 - 1100 Break

1100 - 1200 General Session VII - Thayer Hall, Room 342

**Chairperson: Jagdish Chandra, U.S. Army Research Office,
Research Triangle Park, North Carolina**

VISUALIZATION IN COMPUTATIONAL FLUID DYNAMICS

Paul Woodward, University of Minnesota, Minneapolis, Minnesota

1200 - 1215 ADJOURNMENT

TOWARD A NEW METHOD OF DECODING ALGEBRAIC CODES USING GRÖBNER BASES

A. Brinton Cooper, III
U.S. Army Research Laboratory
Aberdeen Proving Ground, Maryland 21005-5066
USA

Abstract

A binary BCH error control code is a vector subspace of binary n -tuples. Algebraically, the code is generated by a polynomial having binary coefficients and roots in $\text{GF}(2^m)$. It is decoded by computing a set of syndrome equations which are multivariate polynomials over $\text{GF}(2^m)$ and which exhibit a certain symmetry. If the number of transmission errors in a received word does not exceed a bound t for the code, the roots of the syndromes are the locations, in the received word, of those errors. These multivariate polynomials are taken as the basis for an ideal in the ring of polynomials in t variables over $\text{GF}(2^m)$. A celebrated algorithm by Buchberger produces a reduced Gröbner basis of that ideal. It turns out that, since the common roots of all the polynomials in the ideal are a set of isolated points, this reduced Gröbner basis is in triangular form, and the univariate polynomial in that basis is the well-known BCH error locator polynomial, the roots of which specify the error locations. Decoding is algorithmically complete when this polynomial is known.

1 Introduction

Modern algebraic techniques have been used to design and decode codes for error control as far back as the presentation of Hamming's codes [1]. In the early 1960s, the binary BCH¹ codes [3-6] were discovered independently by Bose and Chaudhuri and by Hocquenghem. The BCH codes and their descendants are popular for several reasons, including their regular algebraic structure which permits easy encoding using simple shift registers and the existence of codes for a wide range of block lengths and error correction capabilities.

However, the asymptotic performance of BCH codes is not "good" [6] in that the error probability after decoding and the information rate of the code are not simultaneously bounded away from zero with increasing block length. Nevertheless, the BCH codes and

¹Bose-Chaudhuri-Hocquenghem. McEliece [2] presents an interesting history of the naming of these codes.

their derivatives are widely used because they are easy to generate, well understood, and useful in the control of transmission errors over noisy channels. Decoders, however, are complex, and work continues to find simpler and more powerful decoders.

This work applies recent results from the algebra of multivariate polynomials to the direct solution of the syndrome equations of binary BCH codes. In this problem, up to t nonlinear polynomial equations must be solved for the locations of the errors.

Following a review of the basic theory of linear block codes, Section II presents the polynomial model of cyclic codes and shows how a BCH code is specified solely by a set of roots of its generator polynomial. Section III defines the essential problem for decoding BCH codes. Section IV casts the problem into ideals in the polynomial ring $GF(2^m)[X_1, \dots, X_t]$, such ideals being defined by their roots. Modern methods are used to solve these equations directly.

Examples are included.

2 Linear Block Codes

2.1 Introduction

A common method for controlling errors in information transmitted over noisy channels is the use of *linear block codes* (LBC)². Algebraically, a LBC is a subspace of a vector space of n -tuples over a finite field and, therefore, has a *basis* which spans the code. The dimension of the LBC is smaller than n , the number of elements in the n -tuple. This gives rise to the existence of $n - k$ *redundant symbols* in each codeword. This redundancy provides *distance* between pairs of codewords. The sense in which we define "nearness" is Hamming distance.

Definition: The Hamming distance d_H between two n -tuples is the number of places in which they differ.

Channel noise often reduces the distance between two received words. Sufficient redundancy, however, can provide enough inter-codeword distance to protect against specified levels of channel noise.

²For a thorough coverage of this topic, the reader is referred to any of several excellent texts [2-9].

2.2 Polynomials and Cyclic Codes

Using powers of an indeterminate x as placeholders permits writing a polynomial model of the LBC. This is more than formalism, however, as it permits code construction and decoding based upon the roots of certain polynomials.

Information is carried in the (binary) coefficients of

$$i(x) = i_0 + i_1x + \cdots + i_{k-1}x^{k-1}, \quad i_j \in \text{GF}(2), \quad j = 0, 1, \dots, k-1. \quad (1)$$

Codeword polynomials are generated by multiplying $i(x)$ by a *generator polynomial* $g(x)$ of degree $n - k$:

$$g(x) = g_0 + g_1x + \cdots + g_{n-k}x^{n-k}, \quad g_j \in \text{GF}(2), \quad j = 0, 1, \dots, n-k \quad (2)$$

Coefficients of the resulting polynomial $v(x)$ represent the binary symbols in the codeword:

$$\begin{aligned} v(x) &= i(x)g(x) \\ &= v_0 + v_1x + \cdots + v_{n-1}x^{n-1}, \quad v_j \in \text{GF}(2), \quad j = 0, 1, \dots, n-1 \end{aligned} \quad (3)$$

A code is said to be *cyclic* if every cyclic shift of every codeword is also a codeword. Algebraically, it is true that a code is cyclic whenever $g(x) | x^n - 1$, and it follows that a cyclic code is an ideal³ in the ring of polynomials modulo $x^n - 1$. More important, the codeword length n is the smallest integer for which $g(x) | x^n - 1$.

2.3 BCH Codes

The BCH codes provide a convenient paradigm for several families of powerful LBCs including Reed-Solomon [2-9] and Goppa [2] codes. A binary, primitive BCH code is a cyclic code of length $n = 2^m - 1$. Its generator polynomial numbers among its roots $2t$ consecutive powers⁴ of a primitive element α of the locator field $\text{GF}(2^m)$. With correct decoding, this code can correct up to t channel errors in every codeword.⁵

Example: Let $m = 4$ and $t = 2$. Then $n = 15$ and the roots of $g(x)$ include α , α^2 , α^3 , and α^4 . Because $\alpha^{15} = 1$, these must also be roots of $g(x)$: $\{\alpha^8, \alpha^6, \alpha^{12}, \alpha^9\}$.

³A formal definition of *ideal* is given later.

⁴The nonzero powers $\alpha^0, \alpha^1, \dots, \alpha^{2^m-2}$ of a primitive element of $\text{GF}(2^m)$ are the distinct nonzero elements of that field.

⁵In order that the codewords be binary, it is necessary, for every root β^i of $g(x)$, that all conjugates $\{\beta^{2^i}, \beta^{4^i}, \dots\}$ be roots of $g(x)$ as well.

Hence, the degree of $g(x)$ is $n - k = 8$ so that the dimension k of the code is 7. (i.e., the code has $2^7 = 128$ code words.) The code is capable of correcting at least $t = 2$ errors in every codeword, and the code rate, k/n is 0.47 information bits per binary symbol transmitted.

3 BCH Decoding

Of course correcting t errors in a codeword of length n requires a decoding procedure that can achieve this error correcting potential. A trivial but completely correct decoding technique is to construct a table of every received binary n -tuple and the codeword into which it must be decoded. When symbol errors are independent, the rule is to decode a received n -tuple into the nearest codeword⁶

However, such table lookup decoding is feasible only for rather small codes, so we continue to be interested in algorithmic, algebraic decoders which are much faster and demand much less storage. Let $r(x)$ represent the received vector when a t -error correcting BCH codeword $v(x)$ is transmitted over a channel corrupted by additive noise:

$$r(x) = v(x) + e(x). \quad (4)$$

$e(x)$ is the error polynomial: $e_j = 1$ if an error occurred in the j^{th} position and 0 otherwise. The paradigm for many useful decoders of this code is a four-step decoding procedure:

- calculate *syndromes*, functions of the coefficients of $r(x)$;
- calculate coefficients of the *error locator polynomial*;
- solve the error locator polynomial for the locations of the errors; and
- (for nonbinary codes) calculate the error values.

3.1 The Syndromes

Consider the channel output, $r(x)$ as given by (4). The j^{th} syndrome value is:

$$S_j = r(\alpha^j) = g(\alpha^j) + e(\alpha^j) = e(\alpha^j), \quad j = 1, \dots, 2t \quad (5)$$

⁶Because this is a *minimum distance decoding* technique, no other decoder can correct more errors on a memoryless channel.

Writing only those coefficients e_j which are not zero leads to the following form of the $2t$ syndrome equations:

$$\begin{aligned} e_{i_1}\alpha^{i_1} + e_{i_2}\alpha^{i_2} + \dots + e_{i_t}\alpha^{i_t} &= S_1 \\ e_{i_1}\alpha^{2i_1} + e_{i_2}\alpha^{2i_2} + \dots + e_{i_t}\alpha^{2i_t} &= S_2 \\ &\vdots \\ e_{i_1}\alpha^{2ti_1} + e_{i_2}\alpha^{2ti_2} + \dots + e_{i_t}\alpha^{2ti_t} &= S_t \end{aligned} \quad (6)$$

Note the following:

(a) In (6), if α^{i_j} for any j is known, then the location i_j of the corresponding error also is known. It is convenient, therefore, to write $X_j = \alpha^{i_j}$. The values of the α^{i_j} are called the *error locators* of the received word.

(b) In any field $GF(2^m)$ of characteristic two, $(a+b)^2 = a^2 + b^2$. Therefore, in (6) every syndrome computed from even powers of α is an even power of some syndrome computed from odd powers of α ; e.g., $S_2 = S_1^2$. These are redundant and do not contribute to solving for the error locators.

(c) In (6), $e_{i_j} = 1$, $j = 1, \dots, 2t$. The syndromes $\{S_j, j = 1, \dots, 2t\}$ are known (computed) elements of $GF(2^m)$ and can be expressed as powers of α ; i.e., $S_\sigma = \alpha^{j_\sigma}$.

Considering (a), (b), and (c) with (6) gives a system of t polynomial equations, the solutions to which are the error locators of the received word.

$$\begin{aligned} S_1 &= \alpha^{j_1} = X_1 + X_2 + \dots + X_t \\ S_3 &= \alpha^{j_3} = X_1^3 + X_2^3 + \dots + X_t^3 \\ &\vdots \\ S_{2t-1} &= \alpha^{j_{2t-1}} = X_1^{2t-1} + X_2^{2t-1} + \dots + X_t^{2t-1} \end{aligned} \quad (7)$$

3.2 The Error Locator Polynomial

Derivation of (7), a set of power-sum symmetric functions, assumed that no more than t errors occurred in a block of length n . The error locator polynomial is derived from these functions.

Definition: The error locator polynomial $\sigma(x)$ is the (univariate) polynomial all the roots of which indicate the locations of errors in a received word.

$$\begin{aligned}\sigma(x) &= \prod_{i=1}^t (x - X_i) \\ &= x^t + \sigma_1 x^{t-1} + \sigma_2 x^{t-2} + \dots + \sigma_t\end{aligned}\quad (8)$$

Decoding is complete when the roots of $\sigma(x)$ are found and the necessary corrections made to $r(x)$. The *Chien search* [8] is a method for doing this without explicitly solving $\sigma(x)$. This method uses a digital circuit which evaluates $\sigma(x)$ at each member α^j of $\text{GF}(2^m)$ and sets a *correction bit* to unity if $\sigma(x)$ is satisfied. The received polynomial $r(x)$ is clocked through the circuit and the correction bit is added module 2 at the appropriate location. Whenever a root of $\sigma(x)$ is found, therefore, the appropriate received symbol is complemented. The Chien search will be required in implementing the direct solution methods discussed below.

4 Direct Solution Techniques

The objective is to find a solution set to (7):

$$\begin{aligned}\alpha^{j_1} &= X_1 + X_2 + \dots + X_t \\ \alpha^{j_2} &= X_1^3 + X_2^3 + \dots + X_t^3 \\ &\vdots \\ \alpha^{j_{2t-1}} &= X_1^{2t-1} + X_2^{2t-1} + \dots + X_t^{2t-1}\end{aligned}\quad (9)$$

where α is a primitive element in $\text{GF}(2^m)$. Because the number of errors in a received word does not exceed t , (9) is a system F of t independent equations with at most t solutions. Hence F is a system of t polynomials in t unknowns and has one unique solution, $\beta = (\beta_1, \dots, \beta_t)^T$.

4.1 Rings and Ideals

Direct solution techniques of (9) exploit the rich algebraic structure of the *ring* $R = K[X] = K[X_1, X_2, \dots, X_t]$ of polynomials in t variables over $K = \text{GF}(2^m)$ [11]. A subset \mathcal{I} of a ring is called an *ideal* if it is a subgroup of the additive group of the ring and if, for every $i \in \mathcal{I}$

⁷Actually, the rigorously correct statement is that all zeros of the system are "equivalent" and "mapped on one another by an isomorphism which leaves fixed the elements of the ground field..." [10]

and every $r \in R$, both ir and ri belong to \mathcal{I} . Hilbert's Basis Theorem [10] requires that every ideal in $K[X]$ have a finite basis.

Consider F to be a subset of the ring $K[X]$. The set $\mathcal{I}(F)$ spanned by members of F (where coefficients are taken from $K[X]$) is an ideal in $K[X]$:

$$\mathcal{I}(F) \triangleq (F) \subset K[X]. \quad (10)$$

The common zeros of the polynomials of F are said to form an *algebraic manifold*, [10] which is "defined by" those polynomials. All points of the manifold satisfy every member of $\mathcal{I}(F)$. Direct solution techniques require searching $\mathcal{I}(F)$ for another set G of polynomials which are simpler to solve than those in F . Hence, new methods for finding bases of ideals in $K[X]$ bear on the decoding problem.

4.2 A Basis for the Ideal

The objective is to find for $\mathcal{I}(F)$ a basis G which is "easily" solved for the underlying roots.

The basis G is obtained from the defining polynomial set F by applying transformations which do not eliminate any roots of the system.

Example: Suppose set F is:

$$\begin{aligned} f_1 &: X_1 + X_2 + \alpha^j = 0 \\ f_2 &: X_1^3 + X_2^3 + \alpha^k = 0, \end{aligned} \quad (11)$$

and suppose that it is known that this system has the solution $(\beta_1, \beta_2) \in \text{GF}(2^m)^2$. Then

$$y(X) = a_1(X)f_1(X) + a_2(X)f_2(X) \quad (12)$$

is satisfied by (β_1, β_2) as well⁸.

Suppose $a_2(X) = 1$ and

$$a_1(X) = X_1^2 + X_1(X_2 + \alpha^j) + (X_2 + \alpha^j)^2. \quad (13)$$

Then,

$$y(X) = X_2^2 \alpha^j + X_2 \alpha^{2j} + \alpha^{2j} + \alpha^k, \quad (14)$$

⁸Of course, if $a_1(X)$ and $a_2(X)$ have a common factor, $y(X)$ may have an additional root that does not satisfy f_1 or f_2 , but this case is of no interest.

and the system has been *reduced* from two equations (a cubic and a linear) to a single, univariate, second degree equation having the same solution (β_1, β_2) . We say that the cubic has been *reduced modulo F* to $y(X)$.

The algorithm for deriving the desired ideal basis G is based upon such reduction operations. It produces a *reduced Gröbner basis* [12] of the ideal spanned by F . A reduced Gröbner G basis is a set of polynomials:

- which is a basis of the ideal;
- each member of which has coefficient of highest order term = 1;
- no element of which can be reduced modulo G .

It is known [12] that a reduced Gröbner basis for $\mathcal{I}(F)$: can be written in *triangularized* form:

$$\begin{aligned} g_1 &= g_1(X_1) \\ g_2 &= g_2(X_1, X_2) \\ &\vdots \\ g_t &= g_t(X_1, X_2, \dots, X_t) \end{aligned} \tag{15}$$

This suggests a recursive root finding technique. However the univariate member g_1 of the set is, in fact, the BCH error locator polynomial [13].⁹

4.3 Gröbner Bases as a Basis for Decoding

It would be redundant to include here the general form of Buchberger's algorithm for finding the Gröbner basis of an ideal $\mathcal{I}(F)$. The interested reader should refer to the literature, of which [12] is the most comprehensive source. The method is illustrated in this example:

Example: This is the general form of the problem. Take K to be $\text{GF}(2^4)$, and $t = 3$. Then the resulting $g(x)$ generates a 3-error correcting code with block length $n = 2^4 - 1$,

⁹At worst, g_1 is isomorphic to the error locator polynomial. As shown in [13] however, the isomorphism is trivial.

dimension $k = 5$, and 32 code words. In general, the decoder produces these non-redundant syndromes:

$$\begin{aligned} X_1 + X_2 + X_3 + \alpha^i &= 0 \\ X_1^3 + X_2^3 + X_3^3 + \alpha^j &= 0 \\ X_1^5 + X_2^5 + X_3^5 + \alpha^k &= 0. \end{aligned} \quad (16)$$

Define three intermediate polynomials,

$$\begin{aligned} p_1(X) &= \sum_{j=0}^2 X_3^j (X_2 + X_1 + \alpha^i)^{2-j} \\ p_2(X) &= \sum_{j=0}^4 X_3^j (X_2 + X_1 + \alpha^i)^{4-j} \\ p_3(X) &= X_2^2 + X_2 X_1 + X_2 \alpha^i + X_1^2 + X_1 \alpha^i + \alpha^{2i}, \end{aligned} \quad (17)$$

and from these produce three "coefficient" polynomials:

$$\begin{aligned} a_1(X) &= p_1 p_3 (X_1 + \alpha^i) + p_2 (X_1 + \alpha^i) + p_1 (\alpha^j + \alpha^{3i}) \\ a_2(X) &= p_3 (X_1 + \alpha^i) + \alpha^j + \alpha^{3i} \\ a_3(X) &= X_1 + \alpha^i. \end{aligned} \quad (18)$$

Substitute the p_i into the a_j to get

$$\begin{aligned} a_1(X) &= X_1 X_3^4 + \alpha^i X_3^4 + X_1 X_2 X_3^3 + \alpha^i X_2 X_3^3 + X_1^2 X_3^3 + \alpha^{2i} X_3^3 + X_1^2 X_2 X_3^2 \\ &\quad + \alpha^{2i} X_2 X_3^2 + \alpha^i X_1^2 X_3^2 + \alpha^{2i} X_1 X_3^2 + \alpha^j X_3^2 + \alpha^{3i} X_3^2 + X_1^2 X_2^2 X_3 + \alpha^{2i} X_2^2 X_3 \\ &\quad + X_1^3 X_2 X_3 + \alpha^j X_2 X_3 + \alpha^i X_1^3 X_3 + \alpha^j X_1 X_3 + \alpha^{j+i} X_3 + \alpha^{4i} X_3 + X_1^2 X_2^3 \\ &\quad + \alpha^{2i} X_2^3 + \alpha^i X_1^2 X_2^2 + \alpha^{2i} X_1 X_2^2 + \alpha^j X_2^2 + \alpha^{3i} X_2^2 + X_1^4 X_2 + \alpha^{4i} X_2 \\ &\quad + \alpha^i X_1^4 + \alpha^{2i} X_1^3 + \alpha^j X_1^2 + \alpha^{4i} X_1 + \alpha^{j+2i} + \alpha^{5i} \\ a_2(X) &= X_1 X_2^2 + \alpha^i X_2^2 + X_1^2 X_2 + \alpha^{2i} X_2 + X_1^3 + \alpha^j. \end{aligned}$$

This yields a univariate polynomial which we recognize as the error locator polynomial:

$$\begin{aligned} \sigma(X_3) &= \sum_{\nu=1}^3 a_\nu(X) f_\nu(X) \\ &= X_3^3 (\alpha^j + \alpha^{3i}) + X_3^2 (\alpha^{i+j} + \alpha^{4i}) + X_3 (\alpha^k + \alpha^{2i+j}) \\ &\quad + \alpha^{i+k} + \alpha^{2j} + \alpha^{3i+j} + \alpha^{6i}. \end{aligned} \quad (19)$$

Finding $\sigma(X)$ solves the decoding problem.

5 Conclusion

Mathematically, we have shown a decoder that computes a set of syndrome values which are functions of the roots of the code's generator polynomial and of the error locations. These syndromes are the constant terms of a system of nonlinear polynomials. We have presented a method for extracting from that system the error locator polynomial, which is satisfied by the error locations expressed as elements of $GF(2^m)$. The coefficients of the error locator polynomial are functions of the syndrome values only. Thus, the decoder need do only two things: compute syndromes and coefficients.

Work is ongoing to generalize this method and to extend it to Reed-Solomon and Goppa codes.

References

- [1] Hamming, R.W., "Error detecting and error correcting codes," *B.S.T.J* 29 (April, 1950), 147-160.
- [2] R.J. McEliece, "The Theory of Information and Coding," in *Encyclopedia of Mathematics and its Applications, Volume 3*, Cambridge University Press, Cambridge, 1977.
- [3] W.W. Peterson and E J Weldon, Jr, *Error-Correcting Codes*, MIT Press, Cambridge, 1972.
- [4] Lin, S. & D.J. Costello Jr., *Error Control Coding*, Prentice-Hall, Englewood Cliffs, 1983.
- [5] Blahut, R.E., *Theory and Practice of Error Control Codes*, Addison-Wesley, Reading, 1983.
- [6] MacWilliams, F.J. & N.J.A. Sloane, *The Theory of Error Correcting Codes*, North-Holland, Amsterdam, 1977.
- [7] Berlekamp, E.R., *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

- [8] Michelson, A.M. & A.H. Levesque, *Error-Control Techniques for Digital Communication*, Wiley, New York, 1985.
- [9] Pless, V., *Introduction to the Theory of Error-Correcting Codes*, Wiley-Interscience, New York, 1982.
- [10] van der Waerden, B.L., *Modern Algebra – Volume II*, Frederick Ungar, New York, 1950.
- [11] van der Waerden, B.L., *Modern Algebra – Volume I*, Frederick Ungar, New York, 1953.
- [12] Buchberger, B., "An Algorithmic Method in Polynomial Ideal Theory," in *Multidimensional Systems Theory*, N.K. Bose, ed., Mathematics and Its Applications, D. Reidel, Boston, 1985.
- [13] Cooper, A.Z., III, "Finding BCH error locator polynomials in one step," *Electronics Letters* 27 (24 October 1991), 2090–2091.

An Algorithm for the Computation of Gröbner Bases

W. W. Adams ^{*} A. Boyle [†] P. Loustau [‡]

Abstract

Let R be a Noetherian integral domain which is graded by an ordered group Γ and let \mathbf{x} be a set of n variables with a term order. In this paper we present a new algorithm for computing Gröbner bases in the ring $R[\mathbf{x}]$. This algorithm is based on the authors earlier paper [2]. In the case where $R = k[y]$ is graded by a term order, then this gives a new algorithm for computing Gröbner bases in $k[y, \mathbf{x}]$. This algorithm requires the computation of many Gröbner bases but in fewer variables than the usual Buchberger Algorithm.

1 Introduction

Let y and \mathbf{x} be sets of variables, each with a term order and an elimination order between them. Let k denote a Noetherian commutative ring. In several places in the literature (e.g. [1], [2], [3], [5], [9], [10]), the problem of lifting Gröbner bases from the ring $k[y]$ to the ring $k[y, \mathbf{x}]$ has been investigated. This entails understanding the difference between a Gröbner basis in $(k[y])[x]$ and a Gröbner basis in $k[y, \mathbf{x}]$. We refer to this as the transitivity question. This problem was examined in [1] mainly for the case when $\mathbf{x} = x$ consisted of a single variable. There it was shown that certain subsets of the leading

^{*}University of Maryland, College Park, MD

[†]National Science Foundation, Washington D.C.

[‡]George Mason University, Fairfax, VA

coefficients with respect to x must form Gröbner bases in $k[y]$ in order to go from a Gröbner basis in $(k[y])[x]$ to a Gröbner basis in $k[y, x]$. The results in [1] were generalized in [2] to the case where \mathbf{x} is more than one variable. The key concept was the so-called saturated sets of polynomials as introduced by Möller in [7] (where they were called maximal sets). In this paper we provide a preliminary discussion of how these ideas can be used to give a new algorithm for computing Gröbner bases.

In [2] we found that graded rings are a natural setting for the transitivity question. So, let R be a Noetherian integral domain and assume that R is graded by an ordered group Γ . The concept of Gröbner basis can be extended to such graded rings R (see, for example [2], [6]). If \mathbf{x} denotes a set of n variables, the ring $R[\mathbf{x}]$ can be graded both by $\Gamma \times \mathbb{Z}^n$ and $\{0\} \times \mathbb{Z}^n$. Then the transitivity question in this setting becomes: when is a Gröbner basis in $R[\mathbf{x}]$, graded by $\{0\} \times \mathbb{Z}^n$, also a Gröbner basis in $R[\mathbf{x}]$, graded by $\Gamma \times \mathbb{Z}^n$. The solution is stated in terms of certain sets of leading coefficients, corresponding to the so-called saturated subsets, being Gröbner bases in R .

In Section 2 we will give the definitions for the generalization of the concept of Gröbner bases to graded rings and state the usual characterizations of Gröbner bases in this context. In Section 3 we will give some computability conditions on a graded ring so that we can give the usual Buchberger algorithm for computing Gröbner bases in such rings. In Section 4 we recall from [2] the results on transitivity that allow us to give our new algorithm. This algorithm is presented in Section 5 and an example is given computing a Gröbner basis by this method.

2 Graded Rings

In this section we briefly review the definitions for the theory of Gröbner bases in a general graded ring. For more details see [2].

Let Γ be an additive abelian group which is totally ordered with respect to an order, denoted by " $<$ ", and where we assume that the order respects the group law. The latter means that for all $\gamma, \delta, \eta \in \Gamma$, we have that $\gamma < \delta$ implies that $\gamma + \eta < \delta + \eta$. We assume that R is a Noetherian integral domain graded by Γ . Thus $R = \bigoplus_{\gamma \in \Gamma} R_\gamma$, where each R_γ is an additive abelian group and $R_\gamma R_\delta \subseteq R_{\gamma+\delta}$ for all $\gamma, \delta \in \Gamma$. Let $\Gamma_0 = \{\gamma \in \Gamma \mid R_\gamma \neq 0\}$. Since $0 \in \Gamma_0$ ($1 = 1^2 \in R_0$), we see that Γ_0 is, in fact, a submonoid of Γ . We will assume

that Γ_0 generates Γ . We will also assume that Γ_0 is well ordered. We note that this is equivalent to $0 \leq \gamma$ for all $\gamma \in \Gamma_0$.

Now for each $a \in R$ ($a \neq 0$) we may write $a = \sum_{\gamma \leq \gamma_0} a_\gamma$, with $a_\gamma \in R_\gamma$ and $a_{\gamma_0} \neq 0$. We define $lt(a) = a_{\gamma_0}$ and $v(a) = \gamma_0$. Set $lt(0) = 0$ and $v(0) = 0$. We call $lt(a)$ the *leading term* of a and $v(a)$ the *value* of a . Since R is an integral domain, we have, for all $a, b \in R$, $lt(ab) = lt(a)lt(b)$, and if $ab \neq 0$ then $v(ab) = v(a) + v(b)$. For a subset $F \subseteq R$, we set $Lt(F) = \langle lt(a) | a \in F \rangle$. (Here the symbol $\langle \dots \rangle$ denotes the ideal generated by \dots .) Clearly $Lt(F)$ is a homogeneous ideal.

Definition 2.1 Let I be an ideal in R and let F be a subset of I . We call F a *Gröbner basis* for I provided that F is finite and $Lt(F) = Lt(I)$.

We will say that a subset F of R is a *Gröbner basis* provided that F is a Gröbner basis of the ideal, $\langle F \rangle$, that it generates.

In our context we also have reduction. Let $F = \{f_1, \dots, f_s\}$ be a finite subset of R . For $f, g \in R$, we say that f reduces to g modulo F , denoted $f \xrightarrow{F} g$, provided that $f - g = \sum_{i=1}^s a_i f_i$, where $lt(f - g) = lt(f)$ and $v(lt(f - g)) = \max(v(lt(a_i)lt(f_i)))$. We let \xrightarrow{F}_+ denote the transitive, reflexive closure of \xrightarrow{F} . We say that f is *reduced* provided there is no g such that $f \xrightarrow{F}_+ g$.

The following Theorem provides in our context the usual equivalent conditions for a set to be a Gröbner basis. It parallels exactly the corresponding Theorem 1 in [7] and is proved in exactly the same way. (See [2].)

Theorem 2.2 Let $F = \{f_1, \dots, f_s\}$ be a finite subset of R . Then the following statements are equivalent:

1. F is a Gröbner basis.
2. For every $a \in \langle F \rangle$ we can write

$$a = \sum_{i=1}^s r_i f_i \text{ where } v(a) = \max_{1 \leq i \leq s} (v(r_i f_i)) \quad (1)$$

and $r_i \in R$.

3. Let B be any finite basis of the syzygy module consisting of all sequences (b_1, \dots, b_s) where $b_i \in R$ such that $\sum_{i=1}^s b_i lt(f_i) = 0$ and where we may

assume that for all $(b_1, \dots, b_s) \in B$ we have that there is a $\gamma \in \Gamma$ satisfying $v(b_i \text{lt}(f_i)) = \gamma$ for all i such that $b_i \neq 0$. Then for any sequence $(b_1, \dots, b_s) \in B$ we have that

$$a = \sum_{i=1}^s b_i f_i$$

has a representation as in Equation 1 above.

4. $a \xrightarrow{F} + 0$ for every $a \in \langle F \rangle$.

5. For B as in 3 and for every $(b_1, \dots, b_s) \in B$ we have

$$\sum_{i=1}^s b_i f_i \xrightarrow{F} + 0.$$

3 Gröbner Bases in Graded Rings

In order to compute Gröbner bases in a graded ring R we need to make some computability assumptions about R . We first assume that the grading is effective. By that we mean that given any $r \in R$ we can effectively determine $\text{lt}(r)$. It then follows, since Γ_0 is well ordered, that we can effectively decompose r into its homogeneous components. We further assume that we can do effective computations in homogeneous ideals. That is, we can answer the membership and representation questions for homogeneous ideals. We finally assume that we can determine a basis of the syzygy module of a tuple of homogeneous elements of R (we can then determine a homogeneous basis because of our assumption that the grading is effective). (Note that if R is a polynomial ring over a field $k = R_0$, and Γ is given by a term ordering, then these conditions are equivalent to the usual assumptions that one can solve linear equations inside k .)

Using these assumptions, it is easy to see that the process of reduction described in the last section is also effective. This observation and Theorem 2.2 allow us to give an algorithm for the computation of Gröbner bases in R , which is basically the usual Buchberger algorithm over rings (cf [4] or [7]).

ALGORITHM: GradedGB($\{r_1, \dots, r_s\}, \Gamma$)
INPUT: r_1, \dots, r_s in R .
OUTPUT: A Gröbner basis for $\langle r_1, \dots, r_s \rangle$.
 $G := \{r_1, \dots, r_s\}$
 $F := \emptyset$
While $F \neq G$ **do**
 $F := G$
 Compute a homogeneous basis B_1, \dots, B_t
 of the syzygies of $(lt(r) \mid r \in F)$
 For $i = 1$ **to** t **do**
 Set $B_i = (b_1, \dots, b_s)$
 $\sum_{i=1}^s b_i r_i \xrightarrow{F} r$, **where** r **is reduced**
 If $r \neq 0$ **then** $G := G \cup \{r\}$

We note that in this algorithm, the new elements obtained in the Gröbner basis are effectively expressible in terms of the input polynomials. We will need this for our application in Section 5. We also note that given the hypotheses above on R , we are able to compute Gröbner bases for any ideal in R , and hence we are able to answer the ideal membership question and the syzygy question for arbitrary ideals in R .

As an easy example, we consider $R = k[x, y]$ for a field k and $\Gamma = \mathbb{Z}$, where the grading is given by total degree. Let $f_1 = x^2y + xy^2 + y$, $f_2 = y^3 + x + y$. A homogeneous basis for the syzygy module of $(lt(f_1), lt(f_2))$ is given by $(-y^2, x^2 + xy)$. We compute $-y^2 f_1 + (x^2 + xy) f_2 = -y^3 + x^3 + 2x^2y + xy^2$. Since the latter polynomial cannot be reduced by f_1 and f_2 , we denote it by f_3 and add it to the basis. A homogeneous basis for the syzygy module of $(lt(f_1), lt(f_2), lt(f_3))$ is given by $(x+y, -y, y)$. Since $(x+y)f_1 + (-y)f_2 + yf_3 = 0$, the while loop ends and we have that $\{f_1, f_2, f_3\}$ is a Gröbner basis for the ideal $\langle f_1, f_2 \rangle$.

4 Transitivity for Gröbner Bases.

Let $\mathbf{x} = \{x_1, \dots, x_n\}$ be variables. Assume that we have a term order " $<$ " on \mathbb{Z}^n . (We note that this is just an order on the group \mathbb{Z}^n in the sense given in Section 2.) Given any integral domain R there is the natural \mathbb{Z}^n grading on $R[\mathbf{x}]$ whose non-zero homogeneous summands are indexed precisely by \mathbb{N}^n .

Now assume that R is, in fact, a Γ -graded ring as above. Set $\Lambda = \Gamma \times \mathbb{Z}^n$. Then we can define a Λ -grading on $R[\mathbf{x}]$ where, for $\gamma \in \Gamma$ and $\nu \in \mathbb{Z}^n$ we have

$$R[\mathbf{x}]_{(\gamma, \nu)} = R_\gamma \mathbf{x}^\nu$$

provided that $\nu \in \mathbb{N}^n$ and is $\{0\}$ otherwise. (By \mathbf{x}^ν we mean $x_1^{\nu_1} \cdots x_n^{\nu_n}$, where $\nu = (\nu_1, \dots, \nu_n)$.) We see that $\Lambda_0 = \Gamma_0 \times \mathbb{N}^n$. We will define an order on Λ as follows.

Definition 4.1 *The elimination order on Λ is defined as $(\gamma_1, \nu_1) < (\gamma_2, \nu_2)$ if and only if $\nu_1 < \nu_2$ or $\nu_1 = \nu_2$ and $\gamma_1 < \gamma_2$.*

This generalizes the concept of elimination order that occurs in the literature, for example, in the computer algebra system Macaulay, and also in [1] and [2]. It is easily seen that Definition 3.1 makes Λ into an ordered group satisfying the conditions assumed above.

Now given $a \in R$ we use the notation $lt_\Gamma(a)$ and $v_\Gamma(a)$ to specify the leading term and value of a as defined in the previous section. If $f \in R[\mathbf{x}]$ we denote the same concepts with respect to Λ by $lt_\Lambda(f)$ and $v_\Lambda(f)$ respectively. Write $f = a\mathbf{x}^\nu +$ lower terms in \mathbf{x} , where $a \in R$ and $a \neq 0$. Then set $lt_x(f) = a\mathbf{x}^\nu$, $lp_x(f) = \mathbf{x}^\nu$, $lc_x(f) = a$ and $v_x(f) = \nu$. Of course, lt_x and v_x are the leading term and value concepts in $R[\mathbf{x}]$ with respect to the group $\{0\} \times \Gamma$. Also, $lp_x(f)$ is called the *leading power product* of f and $lc_x(f)$ is called the *leading coefficient* of f . We note that $lt_\Lambda(f) = lt_\Lambda(lt_x(f)) = lt_\Gamma(lc_x(f))lp_x(f)$ and $v_\Lambda(f) = (v_\Gamma(lc_x(f)), v_x(f))$. We also define Lt_Γ , Lt_x and Lt_Λ as in the last section. The former will give homogeneous ideals in R and the latter will give ideals in $R[\mathbf{x}]$, homogeneous with respect to $\{0\} \times \mathbb{Z}^n$ and Λ respectively.

In order to state our algorithm we must relate Gröbner bases in $R[\mathbf{x}]$ with respect to $\Lambda = \Gamma \times \mathbb{Z}^n$ to Gröbner bases in $R[\mathbf{x}]$ with respect $\{0\} \times \mathbb{Z}^n$ and Gröbner bases in R with respect to Γ of certain subsets of leading coefficients. We now define these subsets of R . Let $F = \{f_1, \dots, f_s\}$ be a set of polynomials in $R[\mathbf{x}]$. We adopt the notation that

$$f_i = a_i X_i + \text{lower terms in the } \mathbf{x} \text{ variables,}$$

where X_i is a power product in the \mathbf{x} variables, and $a_i \in R$. That is, $lc_x(f_i) = a_i$ and $lp_x(f_i) = X_i$. We will continue using this notation throughout the paper.

For each subset S of $\{1, \dots, s\}$, we define

- $D_S = \text{lcm}_{i \in S} X_i = \text{lcm}_{i \in S} \{lp_x(f_i)\}$,
- $F_S = \{f_i | X_i \text{ divides } D_S\}$, and
- $G_S = \{a_i | f_i \in F_S\}$.

Also let $S^* = \{i | f_i \in F_S\}$. We say that S is *saturated* if $S^* = S$.

Theorem 4.2 F is a Gröbner basis in $R[x]$ with respect to $\Lambda = \Gamma \times \mathbb{Z}^n$ if and only if

1. F is a Gröbner basis in $R[x]$ with respect to $\{0\} \times \mathbb{Z}^n$ and
2. For all saturated subsets S of $\{1, \dots, s\}$, G_S is a Gröbner basis in R with respect to Γ .

The proof is given in [2].

The next theorem is the basis of the algorithm we will give in Section 5. It is easily deduced from Theorem 4.2 (see [2]). It shows that if we can compute Gröbner bases in $R[x]$ with respect to $\{0\} \times \mathbb{Z}^n$, and in R with respect to Γ , then we can compute Gröbner bases in $R[x]$ with respect to $\Gamma \times \mathbb{Z}^n$.

Theorem 4.3 Let R be a Γ -graded ring, and $F = \{f_1, \dots, f_s\}$ be a Gröbner basis in $R[x]$ with respect to $\{0\} \times \mathbb{Z}^n$. For each saturated subset S of $\{1, \dots, s\}$ let $\{a_{S,1}^*, \dots, a_{S,t_S}^*\}$ be a Gröbner basis of G_S in R with respect to Γ . Write

$$a_{S,i}^* = \sum_{j \in S} b_{S,i,j} a_j$$

and define

$$f_{S,i}^* = \sum_{j \in S} b_{S,i,j} \frac{D_S}{X_j} f_j$$

for all $i = 1, \dots, t_S$. Then

$$F^* = \bigcup_{S \text{ saturated}} \{f_{S,1}^*, \dots, f_{S,t_S}^*\} \cup F$$

is a Gröbner basis with respect to Λ .

The above theorem generalizes a result of Möller in [7]. There $\Gamma = 0$, and R is a PID and the polynomials $f_{S,i}^*$'s are called T-polynomials.

5 A New Algorithm for Computing a Gröbner Basis in $R[x]$

Now based on Theorem 4.3 we give our new algorithm for computing a Gröbner basis for an ideal in $R[x]$ with respect to $\Gamma \times \mathbb{Z}^n$, which differs from the usual one as given in Section 3. The algorithm in Section 3 will be used to compute Gröbner bases in $R[x]$ with respect to $\{0\} \times \mathbb{Z}^n$ and to compute Gröbner bases in R with respect to Γ . (We note that the computability assumptions on R made in Section 3 are inherited by $R[x]$ for either of the gradings.)

INPUT: $g_1, \dots, g_t \in R[x]$

OUTPUT: A Gröbner basis of $\langle g_1, \dots, g_t \rangle$ with respect to $\Gamma \times \mathbb{Z}^n$.

$F = \text{GradedGB}(\{g_1, \dots, g_t\}, \{0\} \times \mathbb{Z}^n)$

Set $F = \{f_1, \dots, f_s\}$ and set $lp_x(f_i) = X_i$ and $lc_x(f_i) = a_i$.

For each saturated subset S of $\{1, \dots, s\}$ do

$G_S^* = \text{GradedGB}(G_S, \Gamma)$

For each $a^* \in G_S^*$ do

Write $a^* = \sum_{i \in S} b_i a_i$

Set $f^* = \sum_{i \in S} b_i \frac{D_S}{X_i} f_i$

$F = F \cup \{f^*\}$

If we specialize to the case where $R = k[y]$ and the grading on $k[y]$ is given by a term order, then the algorithm above gives a new algorithm for computing a Gröbner basis with respect to an elimination order in the polynomial ring $k[y, x]$. So we can compute a Gröbner basis for a polynomial ring in $m + n$ variables by computing many Gröbner bases in m variables and one in n variables. The algorithm for computing the one basis in n variables relies on being able to compute syzygies over $k[y]$ which may be done by computing Gröbner bases in m variables. (Alternatively, we could use Möller's method of inductively computing the syzygies, cf [7].) We could, of course, apply this same idea to compute the Gröbner bases in m variables obtaining a recursive procedure. Since we are always computing with a smaller number of variables, there is the hope that time could be saved in this manner. However, an initial investigation with a naive implementation on Maple was not too encouraging.

We now give a simple example of computing a Gröbner basis using our algorithm. We consider the ring $\mathbb{Q}[x, y, z, t]$, where \mathbb{Q} is the field of rational numbers, with a lexicographic term ordering with $x > y > z > t$. Let $f_1 = xt - xz^2 + yz$, and $f_2 = xyz t^2 + xy + y^2$. We use our algorithm to compute a Gröbner basis for the ideal $I = \langle f_1, f_2 \rangle$. Let $R = \mathbb{Q}[z, t]$, where $\Gamma = \mathbb{Z}^2$ with the ordering defined by lex with $z > t$. We adjoin the variables x and y with the lex ordering with $x > y$. The first step in our algorithm is to compute a Gröbner basis of I with respect to $\{0\} \times \mathbb{Z}^2$ using the algorithm of Section 3. We first need to compute a basis of the syzygy module of $lt(f_1) = x(t - z^2)$, $lt(f_2) = xy(z t^2 + 1)$. This is trivially seen to be $(y(z t^2 + 1), -(t - z^2))$. The corresponding S-polynomial is: $y(z t^2 + 1)f_1 - (t - z^2)f_2 = y^2(z^2 t^2 + z^2 + z - t)$. This polynomial is not reducible, so we call it f_3 . Now we need to compute the syzygies for $lt(f_1), lt(f_2), lt(f_3) = f_3$, but it is easily seen that the corresponding S-polynomials reduce to 0. The next step in our algorithm is to compute the saturated sets where $lp(f_1) = x$, $lp(f_2) = xy$, $lp(f_3) = y^2$, which gives $S_1 = \{1, 2\}$ for $D_{S_1} = xy$ and $S_2 = \{1, 2, 3\}$ for $D_{S_2} = xy^2$. $\langle G_{S_1} \rangle = \langle t - z^2, z t^2 + 1 \rangle$ has Gröbner basis $\{z + t^3, t^5 - 1\}$. Since $z + t^3 = t^2(t - z^2) + z(z t^2 + 1)$, and $t^5 - 1 = t^4(t - z^2) + (z t^2 - 1)(z t^2 + 1)$, we need to add the T-polynomials $f_4 = y t^2 f_1 + z f_2 = xyz + x y t^3 + y^2 z t^2 + y^2 z$ and $f_5 = t^4 y f_1 + (z t^2 - 1) f_2 = x y t^5 - x y + y^2 z t^4 + y^2 z t^2 - y^2$. Finally we note that $\langle G_{S_2} \rangle = \langle t - z^2, z t^2 + 1, z^2 t^2 + z^2 + z - t \rangle = \langle G_{S_1} \rangle$, so no new T-polynomials are needed. Therefore $\{f_1, f_2, f_3, f_4, f_5\}$ is the desired Gröbner basis of I . We note that if f_2 is eliminated the remaining polynomials form a reduced Gröbner basis of I .

As a final comment, we note that there is another concept of Gröbner basis, the so-called strong Gröbner basis (as opposed to the concept we used so far in this paper which is sometimes referred to as a weak Gröbner basis). Namely, if R is a Unique Factorization Domain, and I is an ideal of R , then f_1, \dots, f_s is a strong Gröbner basis for I if for every $f \in I$, there exists a j such that $lt(f_j)$ divides $lt(f)$. Using material in [2] we could have developed an algorithm for this case similar to the one above.

References

- [1] W. W. Adams and A. Boyle, "Some Results on Gröbner basis over Commutative Rings," *Journal of Symbolic Computation*, to appear.
- [2] W. W. Adams, A. Boyle and P. Loustau, "Transitivity for Weak and Strong

- Gröbner Bases*," submitted to Journal of Symbolic Computation.
- [3] D. Bayer and M. Stillman, "A Theorem on Refining Division Orders by the Reverse Lexicographic Order," *Duke Mathematical Journal*, **55** (1987), pp. 321-328.
 - [4] B. Buchberger, "Gröbner Bases: An Algorithmic Method in Polynomial Ideal Theory", "Multidimensional Systems Theory" (N.K.Bose, ed.), D. Reidel Publishing Co., pp. 184-232.
 - [5] P. Gianni, B. Trager and G. Zacharias, "Gröbner basis and Primary Decomposition of Polynomial Ideals," *Journal of Symbolic Computation*, **6** (1988), pp. 148-168.
 - [6] A. Miola and T. Mora, "Constructive Lifting in Graded Structures: A Unified View of Buchberger and Hensel Methods," *Journal of Symbolic Computation*, **6** (1988), pp. 306-323.
 - [7] H. Möller, "On the Construction of Gröbner basis Using Syzygies," *Journal of Symbolic Computation*, **6** (1988), pp. 345-360.
 - [8] L. Robbiano, "Gröbner Bases: A Foundation for Commutative Algebra", preprint.
 - [9] D. Spear, "A Constructive Approach to Commutative Ring Theory," *Proceedings 1977 MACSYMA User's Conference*, pp. 369-376.
 - [10] R. Shtokhamer, "Lifting canonical algorithms from a ring R to the ring $R[x]$," *Journal of Symbolic Computation*, **6** (1988), pp. 169-179.
 - [11] G. Zacharias, "Generalized Gröbner Bases in Commutative Polynomial Rings," Thesis at MIT, dept. of Computer Science.

Long Time Behavior of a Numerical Approximation to a Nonlinear Evolution Problem in Viscoelasticity

Donald A. French *

Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH 45221-0025

June 15, 1992

Abstract

We summarize our results on the analysis of the long time behavior of a numerical approximation of a nonlinear evolution problem which are given in detail in [7] and [8]. The time step scheme is derived using finite elements and is called the continuous time Galerkin (CTG) method. It is implicit, of arbitrary order, and closely related to the implicit Runge Kutta (IRK) methods which are derived from Gauss Legendre integration formulas. The main Theorem of this short note states that the approximate solution to the evolution problem converges to a discrete steady state solution. This behavior is qualitatively correct since the true solutions of the evolution problem also tend to static solutions.

Presented at the *Tenth Army Conference of Applied Mathematics and Computing*, West Point, New York, 16-19 June 1992.

1. Nonlinear evolution problem from viscoelasticity:

We consider a model equation for the one-dimensional motion of a viscoelastic bar which may undergo phase changes. We search for $U = U(x, t)$ which is the displacement of a particle at time t having position x in some reference configuration and satisfies

$$\begin{aligned} U_{tt} &= (\sigma(U_x) + U_{xt})_x \text{ in } (0, 1) \times (0, \infty) \\ U(0, t) &= 0, \text{ for } t \geq 0 \\ (\sigma(U_x) + U_{xt})(1, t) &= 0, \text{ for } t \geq 0 \\ U(x, 0) &= U_0(x) \text{ and } U_t(x, 0) = V_0(x) \text{ in } (0, 1) \end{aligned} \tag{1}$$

*Partially supported by the Army Research Office thru grant 28535-MA.

where

$$\sigma(s) = \begin{cases} s + 1 & \text{if } s \leq -1/2 \\ -s & \text{if } -1/2 < s < 1/2 \\ s - 1 & \text{if } s \geq 1/2. \end{cases}$$

The stress σ is nonmonotone allowing for two phases.

Multiplying the equation by U_t , integrating in x and t , and using the boundary conditions with integration by parts gives the energy equation

$$\mathcal{E}(t_2) + \int_0^1 \int_{t_1}^{t_2} U_{xt}^2 dt dx = \mathcal{E}(t_1) \quad (2)$$

where

$$\mathcal{E}(t) = \int_0^1 \left(\frac{1}{2} U_t^2 + \psi(U_x) \right) dx$$

and ψ is a double-well potential

$$\psi(s) = \int_0^s \sigma(r) dr + \frac{1}{4} \geq 0$$

and $0 \leq t_1 < t_2$. This energy relation is crucial to the analysis of the long-time behaviour of U . The steady state equation is

$$(\sigma(U_x))_x = 0 \quad \text{on } (0, 1) \quad (3)$$

$$U(0) = \sigma(U_x(1)) = 0.$$

Functions U that satisfy the boundary conditions and have derivative, U_x , that takes on the values ± 1 or 0 are weak solutions of (3).

Pego [12] summarizes previous results on (1) and similar equations including studies done on the asymptotic behavior of U as $t \rightarrow \infty$. In the introduction he discusses the results of Dafermos [5], Andrews and Ball [2] and his own. Briefly, they state that problem (1) has a unique global weak solution, $U_t \rightarrow 0$, $U \rightarrow U^\infty$, and $\sigma(U_x^\infty) = 0$.

2. Continuous time Galerkin methods:

In this section we introduce the CTG scheme for problem (1). For simplicity of presentation we discretize space and time with uniform meshes.

$$0 = x_0 < x_1 < \dots < x_M = 1, \text{ and } t_0 < t_1 < t_2 < \dots$$

where $I_n = (x_{n-1}, x_n)$, $h = x_n - x_{n-1}$, $J_n = (t_{n-1}, t_n)$, $k = t_n - t_{n-1}$, and $Q_n = (0, 1) \times J_n$.

The finite element space for the variable x is

$$S_h = \{\chi \in C([0, 1]) : \chi|_{I_n} \in P_1(I_n), n = 1, \dots, M \text{ with } \chi(0) = 0\}.$$

$P_q(J)$ is the set of all polynomials of degree $\leq q$ on the interval J . For the variable t we will use the space

$$V_k = \{\tau \in C([0, \infty)) : \tau|_{I_n} \in P_r(I_n), n = 1, \dots, \infty\}$$

where r is also a positive integer. Letting $V_{hk} = S_h \otimes V_k$ our method is as follows:

$$\left\{ \begin{array}{l} \text{Find } (u, v) \in V_{hk} \times V_{hk} \text{ such that} \\ ((u_t - v, \chi))_n = 0 \quad \forall \chi \in S_h \otimes P_{r-1}(J_n) \\ ((v_t, \lambda))_n + ((u_{xt} + \sigma(u_x), \lambda_x))_n = 0 \quad \forall \lambda \in S_h \otimes P_{r-1}(J_n) \\ \text{where } n = 0, 1, 2, 3, \dots, u(\cdot, 0) = u_0 \cong U_0 \text{ and } v(\cdot, 0) = v_0 \cong V_0. \end{array} \right. \quad (4)$$

The inner product is

$$((v, w))_n = \int_{J_n} \int_0^1 v w dx dt.$$

Letting $\chi = v_t$ and $\lambda = u_t$ we obtain a discrete version of (2)

$$\mathcal{E}^n + \int_{J_n} \int_0^1 u_{xt}^2 dx dt = \mathcal{E}^{n-1} \quad (5)$$

where the sum of the kinetic and stored strain energy at time t_n is

$$\mathcal{E}^n = \int_0^1 \left(\frac{1}{2} v(\cdot, t_n)^2 + v(u_x(\cdot, t_n)) \right) dx.$$

CTG methods were first introduced in the context of ordinary differential equations by Hulme [10]. They are closely related to IRK schemes based on Gauss-Legendre quadrature rules (See [10]). In most applications we expect a k^{2r} rate of convergence for the IRK and CTG schemes. See Akrivis and Dougalis [1]; Baker, Dougalis and Karakasian [4]; and McKinney [11] for more on the IRK schemes. See [9] for more references and discussion of energy preserving schemes.

The following result is proved in [8]:

THEOREM 1: Problem (4) has a unique solution for $k < 1$.

3. Long time behavior:

The key to our analysis of $u(\cdot, t)$ as $t \rightarrow \infty$ is the following Theorem which was originally stated and proved in Elliott [6].

Let

$$Z_h = \{z \in S_h : (\sigma(z_x), \phi_x) = 0 \quad \forall \phi \in S_h\}.$$

$u^n = u(\cdot, t_n)$, and $v^n = v(\cdot, t_n)$.

THEOREM 2: Suppose

$$|||(u^n, v^n)|||^2 = \|u_x^n\|_{L^2(0,1)}^2 + \|v^n\|_{L^2(0,1)}^2 \leq K, \quad (6)$$

$$\lim_{n \rightarrow \infty} (\sigma(u_x^n), \phi_x) = 0 \quad \forall \phi \in S_h, \quad (7)$$

$$\lim_{n \rightarrow \infty} (v^n, \chi) = 0 \quad \forall \chi \in S_h. \quad (8)$$

$$\|(z_1 - z_2)_x\|_{L^2(0,1)} \geq \Delta > 0 \quad \forall z_1, z_2 \in Z_h, \quad (9)$$

and for each $z \in Z_h$,

$$|||(u^n - z, v^n)||| \leq \mu |||(u^{n-1} - z, v^{n-1})||| \quad (10)$$

where $\mu > 1$ is a constant. Then

$$v^n \rightarrow 0 \text{ as } n \rightarrow \infty, \quad (11)$$

$$u^n \rightarrow z \in Z_h \text{ as } n \rightarrow \infty. \quad (12)$$

A complete proof of this Theorem is given in [8]. The verification of the hypotheses (6) - (10) so that the desired results (11) - (12) is also given in [8] and is summarized here. The boundedness property (6) follows from the discrete energy inequality (5). The fact that $u_{xt} \rightarrow 0$ also follows from (5) since

$$\int_0^\infty \int_0^1 u_{xt}^2 dx dt \leq E(0).$$

From this one can prove (7) and (8). The space of steady state solutions can be identified explicitly,

$$Z_h = \{z \in S_h : z_x|_{I_i} \in \{-1, 0, 1\}\}$$

and then (9) follows. Property (10) follows by an energy argument.

4. Conclusions:

By using finite elements to discretize time for a nonlinear evolution problem from viscoelasticity a numerical method which preserves (discretely) the energy of the original problem is obtained. With this property it is shown that the numerical approximations tend to discrete steady states as $t \rightarrow \infty$ just as the true solutions to the evolution problem tend to continuous steady states.

References

- [1] G.D. Akrivis and V.A. Dougalis, "On a class of conservative, highly accurate Galerkin methods for the Schrödinger equation" (Preprint).
- [2] G. Andrews and J.M. Ball, "Asymptotic Behaviour and Changes in Phase in One-Dimensional Nonlinear Viscoelasticity", *J. Differential Equations* 44 (1982) 306-341.
- [3] A.K. Aziz and P. Monk, "Continuous finite elements in space and time for the heat equation", *Math. Comp.* 52 (1989) 255-274.
- [4] J. L. Bona, V. A. Dougalis, and O. A. Karakashian, "Fully Discrete Methods for the Kortewig-de Vries Equation", *Comp. Maths. with Appls.* 12A (1986), 859-884.
- [5] C.M. Dafermos "The mixed initial-value problem for the equations of nonlinear one dimensional viscoelasticity", *J. Diff. Eqns.* 6 (1969), 71-86.
- [6] C. M. Elliott "The Cahn-Hilliard model for the kinetics of phase separation" in *Math. Models for Phase Change Problems* ed. J. F. Rodrigues, Birkhäuser Verlag, 1989.
- [7] D. A. French and S. Jensen: "Behaviour in the large of numerical solutions to one-dimensional nonlinear viscoelasticity by Continuous Time Galerkin methods", *Comp. Meth. Appl. Mech. Eng.* 86 (1991), 105-124.
- [8] D. A. French and S. Jensen: "Long term behaviour of arbitrary order continuous time Galerkin methods for some one-dimensional phase transition problems", (Submitted to *IMA J. Num. Anal.*).
- [9] D. A. French and J. W. Schaeffer "Continuous finite element methods which preserve energy properties for nonlinear problems", *Appl. Math. Comp.* 39 (1991), 271-295.

- [10] B.L. Hulme, "One-step piecewise polynomial Galerkin methods for initial value problems", Math. Comp. 26 (1972), 415-426; see also *ibid* 881-891.
- [11] W. McKinney. "Optimal Error Estimates for High Order Runge-Kutta Methods applied to Evolutionary Equations", Thesis 1989, University of Tennessee.
- [12] R.L. Pego, "Phase Transitions in On Dimensional Nonlinear Viscoelasticity: Admissibility and Stability", Arch. Rat. Mech. 97 (1987), 353-394.

CONDENSATION OF THREE-DIMENSIONAL FINITE ELEMENTS TO SOLVE PROBLEMS OF WAVE PROPAGATION

David W. Sykora¹ and Jose M. Roeset²

SPONSOR: US Army Engineer Waterways Experiment Station
Vicksburg, Mississippi

ABSTRACT

The solution of some three-dimensional (3-D) wave propagation problems can be achieved effectively using a two-dimensional (2-D) finite element formulation not involving assumptions of plane strain. This formulation has been adopted into a new computer code to rapidly solve a broader class of wave propagation problems using 2-D methods without loss of accuracy thus resulting in appreciable savings in computer time.

The formulation of the solution method involves two primary components: the condensation of the dynamic stiffness matrices to produce an equivalent 2-D system and the representation of the distribution of loads in the out-of-plane direction using a Fourier expansion. The 2-D mesh is solved in the frequency and wave-number domain and then inverse Fourier transforms are performed to obtain dynamic displacements at any location.

The purpose of this presentation is to describe the general formulation of the solution method, present some results of validation studies and parametric analyses, and make comparisons of computational effort on the US Army CRAY Y-MP between the new 2-D code and a commercial 3-D solution package.

INTRODUCTION

This paper summarizes a method to estimate the variation of displacements in space and time produced by dynamic loads in complex isotropic media, consisting of dipping, discontinuous, and/or irregular layers, using a numerical approximation method (Sykora and Roeset, 1992). The distinguishing feature of this work is that the formulation allows three-dimensional (3-D) problems to be solved using a two-dimensional (2-D) numerical model. To implement this method, the stratigraphy and material properties of the model cannot vary in a horizontal direction (2-D stratigraphy). However, the distribution and extent of loads may vary in both horizontal directions (3-D load) providing for the analysis of synthetic vibratory sources such as a Vibroseis truck. These types of problems cannot be solved analytically but normally would be solved using a laborious 3-D numerical approximation.

The finite element method was selected for computational solution to permit discretization of geologic systems with numerous materials of arbitrary geometry. The formulation involves two primary components: the condensation

¹ Research Civil Engineer, Geotechnical Laboratory, US Army Engineer Waterways Experiment Station, Vicksburg, MS, 39180.

² Professor, Department of Civil Engineering, University of Texas, Austin, TX, 78712.

of 3-D dynamic stiffness matrices to equivalent 2-D matrices and the representation of the distribution of loads in the out-of-plane direction using a Fourier expansion. This strategy was explicitly proposed for axis-symmetric problems by Winnicki and Zienkiewicz (1979) and Lai and Booker (1991) and for 3-D formulations by Runesson and Booker (1982, 1983) and Lin and Tassoulas (1987). This strategy was used specifically for wave propagation studies in horizontally layered pavement systems by Kang (1990) and Hanazato et al. (1991). The 2-D system of equations are first solved in the frequency and wave-number domain; inverse Fourier transforms are then performed to obtain the solution as a function of out-of-plane distance and time, if so desired.

One objective of this study is to examine the potential for determining elastic moduli of materials in complex systems of soil, rock, and structural materials from measured motions. The Spectral-Analysis-of-Surface-Waves (SASW) method (Nazarian and Stokoe 1985a, 1985b) is an existing method of field measurement and mathematical inversion to determine the moduli of horizontally layered systems. This method involves the use of signal processing techniques on two measured vertical components of motion spaced at equal increments from the vibratory source. A similar procedure of determination is desired for more complex systems. In addition, the use of artificial neural networks holds promise to improve inversion schemes (Rix and Leipski, 1991). Therefore, Rayleigh wave propagation will be of primary interest. Rayleigh waves normally contain most of the energy of wave propagation for the near surface regime and Rayleigh wave energy will attenuate with distance at a much lower rate than body waves. The response at the ground surface is normally of greatest interest since it provides the easiest access for measurements.

Assumptions for Two-Dimensional Systems

A common assumption used to reduce the computational effort for the engineering analysis of stress and strain in boundary value problems of interest for geotechnical engineering applications is that of plane strain. Plane strain implies that the displacements in the direction perpendicular to a two-dimensional plane are equal to zero (Love, 1944). This assumption reduces the analysis of a problem from 3-D to 2-D. Conditions of plane strain require 2-D geometry and boundary conditions and loads that are uniform in the direction perpendicular to the plane under consideration (Timoshenko and Goodier, 1970). A plane wave with particle motion only in the 2-D analysis plane is consistent with this assumption.

A surface load distributed over a finite area induces stresses that vary in three principal directions. If stresses vary in a direction perpendicular to the analysis plane, displacements and strains will be non-zero. Therefore, 3-D loads are inconsistent with plane strain assumptions. This study deals with the analysis of "planar" geosystems which proves to be beneficial from a computational standpoint. The primary assumptions are that the geometry and boundary conditions of the system and the distribution of material properties are planar (2-D) but the loads are non-planar (3-D). This set of conditions has a broader range of applications than that for plane strain while circumventing expensive 3-D solution methods.

PREVIOUS STUDIES

The evolution of the state of knowledge for dynamic loads acting on elastic media, particularly that involving Rayleigh waves, was reviewed to provide insight into which problems have been solved, what approaches were used, what conclusions have been reached, and which studies provide a proper basis for validation of the present formulation. Despite a large number of papers on the subject, few are considered to be useful for comparative purposes with this study because:

- a. Almost all studies consider plane wave propagation.
- b. Many of the studies consider only R-wave energy (do not include in-plane P-SV waves),
- c. Experimental studies generally focused on "thin plate" tests which have plane stress boundary conditions, which are generally incompatible with assumptions for this study, and
- d. Plane strain conditions are generally assumed for theoretical and numerical studies.

The exact solution used for validation of the computer program were published by Kausel (1981) for point, disk, or ring loads acting on axi-symmetric systems. These solutions were derived using discrete Green's functions evaluated numerically and are excellent approximations.

Soil dynamics studies conducted in the 1950's and 1960's using finite difference and finite element methods, and in the 1970's and 1980's using Green's functions and boundary element models, generally assumed plane harmonic waves and horizontally layered media extending to infinity. The subsurface distribution of materials at most sites is not simple nor is it conducive to analytical closed-form solutions of wave propagation problems. Sloping strata of finite length, an irregular ground surface, and/or two-dimensional load distributions are prevalent. The present study describes a procedure to analyze wave propagation in these more complex systems while not assuming plane strain assumptions.

MATHEMATICAL FORMULATION

The mathematical formulation is based on simple principles of Elastodynamics, superposition, Fourier series expansion, and numerical discretization and solution procedures using the finite element method. The set of assumptions is intended to be small, to broaden the class of problems that can be solved. The primary assumption required for the condensation method described herein is that the geometry of the system and material properties are planar (do not vary in some horizontal direction). A number of other assumptions were used to derive the first generation computer code, vib3:

- a. Media are isotropic,
- b. Hysteretic behavior is represented by complex moduli relation,
- c. Source produces vertical, steady-state excitation at one frequency,
- d. Base is rigid, and
- e. Distribution of loads is symmetric about y-axis.

These assumptions are not necessary and some will be phased out in future versions of the code. In addition, the computer code does not allow for transmitting boundaries in the 2-D analysis plane. Rather, the domain must be discretized to include enough area for the motions to attenuate sufficiently before being reflected back to the area of interest.

Field Equations

Two primary sets of variables adequately describe the effect of forces acting on linear systems -- stresses and displacements. These variables exist in the following field equations: stress equilibrium, strain-displacement, and constitutive equations. These three sets of equations are combined in terms of displacements to derive the governing equations for the problem. Wave propagation involves the effects of inertia and deformation of the media. The effects of inertia result from masses being accelerated. The derivations below apply to isotropic materials.

Stress equilibrium equations. The summation of stresses acting on a small rectangular parallelepiped in three-dimensional Cartesian space and Newton's second law of motion neglecting body forces are used to derive the stress equilibrium equations. The equations of motion using indicial notation and the soil mechanics convention of compressive forces as positive and accounting for the symmetry of the Cauchy stress tensor are:

$$\sigma_{ij,j} = -\rho \ddot{u}_i \quad (1)$$

where

$$\begin{aligned} \sigma &= \text{stress components [F/L}^2\text{]} \\ \rho &= \text{mass density [F-s}^2\text{/L]} \\ \ddot{} &= \partial^2 / \partial t^2 \text{ [1/s}^2\text{]} \end{aligned}$$

Strain-displacement equations. The strain-displacement (compatibility) equations are derived from small strain theory. The equations for a displacement field, u , are:

$$\epsilon_{ij} = \frac{1}{2} (u_{i,j} + u_{j,i}) \quad (2)$$

and are often referred to as engineering measures of strain.

Constitutive equations. The constitutive equations provide the means to relate stress and strain; they define the deformability of the material. Individual material layers are assumed to be homogeneous, isotropic, and visco-elastic. To begin the formulation of constitutive relations, consider the simplest case of linear elasticity proposed by Hooke. For homogeneous and isotropic conditions:

$$\sigma_{ij} = \left(\frac{2\nu G}{1-2\nu} \right) \epsilon_{kk} \delta_{ij} + 2G \epsilon_{ij} \quad (3)$$

where

$$\begin{aligned} G &= \text{shear modulus} \\ \nu &= \text{Poisson's ratio, and} \\ \delta_{ij} &= \text{the Kronecker delta: } \begin{cases} \delta_{ij}=0 & \text{if } i \neq j \\ \delta_{ij}=1 & \text{if } i=j \end{cases} \end{aligned}$$

Soil is an inelastic material -- energy dissipates from friction as waves travel through it. This phenomenon is called material damping and mathematical models are used to approximate it in governing equations. One form of damping, called hysteretic, is independent of the frequency of excitation. This form of damping can be introduced into the formulation for frequency-domain analyses through the Correspondence Principle (Wolf 1985). This principle states that the elastic stiffness (in this case shear modulus) is replaced by a complex stiffness to obtain the damped solution. The results of this study are expected to be applied at distances greater than one wavelength from the source (e.g., Nazarian and Stokoe, 1985a; Kang, 1990) where shear strains from synthetic sources are small. The following relationship is commonly used to model linear-hysteretic behavior for small shear strains (and small values of damping):

$$G^* = G(1 + 2i\beta) \quad (4)$$

where

G^* is complex shear modulus

β is the damping ratio [-]

$i = \sqrt{-1}$

The magnitude of damping is considered to be independent of strain (Hardin and Drnevich 1972; Johnston, Toksöz, and Timur 1979; and Toksöz, Johnston, Timur 1979) for the levels of shear strains expected.

Equations of Equilibrium

The three sets of field equations are combined to obtain the governing equations. A stiffness formulation was chosen, that is, a relation in terms of displacements (also referred to as displacement approach). These equations are associated with Navier and can be derived by substituting the strain-displacement equations into the constitutive equations, then, substituting the resulting equations into the equilibrium equations. Assuming that the body forces are zero and applying Newton's second law, the result is:

$$G^* \left[\left(\frac{1}{1-2\nu} \right) u_{j,jj} + u_{i,jj} \right] = -\rho \ddot{u}_i \quad (5)$$

These are the partial differential equations that govern wave propagation in three-dimensional Cartesian space for homogeneous, isotropic materials with no body forces. The partial differential equation is classified as hyperbolic leading to an initial value problem.

Finite Element Method in Three-Dimensional Cartesian Space

The finite element method is a numerical analysis technique used to approximate the response of a continuous body by dividing the domain of interest into a discrete number of subdomains. Boundary conditions and external forces are imposed at discrete nodes where the displacements are calculated. Results can be obtained at any point in the body through the use of interpolation functions. In general, as the subdomains become smaller, the solution converges to that of the continuum. Many textbooks describing the finite element method are available with different sets of notation. The notation below closely follows that used by Zienkiewicz and Taylor (1989).

There are two basic approaches to formulating a problem using the finite element method: the (direct) displacement method and the variational method. The displacement method is the most popular and most easily understood procedure (Zienkiewicz and Taylor, 1989) and was selected for this study. The displacement method can be easily used with Fourier superposition analysis in the frequency domain for the solution of elastodynamic problems.

Displacement method. Displacements are specified as the unknowns for the displacement method. Letting u represent the vector of displacements at any point and U the vector of displacements at the nodes of a finite element:

$$u = N U \quad (6)$$

where N is the matrix of interpolation functions. The strains at any point can be represented as:

$$\epsilon = E u \quad (7)$$

where

$$\epsilon = \begin{Bmatrix} \epsilon_x \\ \epsilon_y \\ \epsilon_z \\ \gamma_{xy} \\ \gamma_{yz} \\ \gamma_{xz} \end{Bmatrix} \quad (8)$$

and

$$E = \begin{bmatrix} \frac{\partial}{\partial x} & 0 & 0 \\ 0 & \frac{\partial}{\partial y} & 0 \\ 0 & 0 & \frac{\partial}{\partial z} \\ \frac{\partial}{\partial y} & \frac{\partial}{\partial x} & 0 \\ 0 & \frac{\partial}{\partial z} & \frac{\partial}{\partial y} \\ \frac{\partial}{\partial z} & 0 & \frac{\partial}{\partial x} \end{bmatrix} \quad (9)$$

Then,

$$\epsilon = E u = E N U = B U \quad (10)$$

where B is a matrix containing the corresponding derivatives of the interpolation functions.

The Correspondence Principle allows the constitutive model to represent hysteretic behavior using complex moduli for solutions in the frequency domain. Superposition is valid because of this linear representation. A frequency domain solution implies that the excitation function must be

periodic. Calling D the complex constitutive matrix of the material:

$$D = \frac{2G^*}{1-2\nu} \begin{bmatrix} 1-\nu & \nu & \nu & 0 & 0 & 0 \\ \nu & 1-\nu & \nu & 0 & 0 & 0 \\ \nu & \nu & 1-\nu & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1-2\nu}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1-2\nu}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1-2\nu}{2} \end{bmatrix} \quad (11)$$

The stress vector at any point is:

$$\sigma = D \varepsilon \quad (12)$$

with:

$$\sigma = \begin{Bmatrix} \sigma_x \\ \sigma_y \\ \sigma_z \\ \tau_{xy} \\ \tau_{yz} \\ \tau_{xz} \end{Bmatrix} \quad (13)$$

Applying the principle of virtual work and making use of the above relations, the equations of motion become:

$$M \ddot{U} + K U = P \quad (14)$$

where M is the mass density matrix defined by:

$$M = \int_V \rho N^T N dv \quad (15)$$

where

ρ = mass density

and K is the (static) stiffness matrix defined by:

$$K = \int_V B^T D B dv \quad (16)$$

The relationships for nodal acceleration, \ddot{U} , and displacement, U , are derived by imposing the steady state condition. Considering the load vector:

$$P = \bar{P} e^{i\omega t} \quad (17)$$

where

\bar{P} = vector of amplitudes of nodal forces
 ω = frequency of excitation (rads/sec)

Then the displacement vector, U , can be written as:

$$U = \bar{U} e^{i\omega t} \quad (18)$$

where

\bar{U} = vector of amplitudes of nodal displacements

and the velocity and acceleration vectors are:

$$\dot{U} = i \omega \bar{U} e^{i\omega t} \quad (19)$$

$$\ddot{U} = -\omega^2 \bar{U} e^{i\omega t} \quad (20)$$

By substituting Equations 18 and 20 into Equation 14 and canceling the exponential term, the equations of motion are:

$$(K - \omega^2 M) \bar{U} = \bar{S} \bar{U} = \bar{P} \quad (21)$$

where \bar{S} is the dynamic stiffness matrix of the system defined by:

$$\bar{S} = K - \omega^2 M \quad (22)$$

The dynamic stiffness matrix is complex and a function of frequency. Equation 21 can be solved using matrix operations incorporated in various solution algorithms ("solvers").

The formulation to this point is specific to steady-state, frequency-domain analyses for homogeneous and isotropic materials. The formulation is applicable to analyses in one, two, and three dimensions and any element configuration. Henceforth, the formulation will be specific to the remaining assumptions and requirements of this study.

3-D finite element. A 3-D, isoparametric, finite element with 16 nodes was chosen to implement the condensation formulation described in the next section. Each node has three degrees of freedom. The element has quadratic interpolation in the analysis plane (x and z) and linear interpolation in the out-of-plane direction (y). Equations 15 and 16 can now be stated in more specific terms using the transformed space:

$$M = \int_{-1}^1 \int_{-1}^1 \int_{-1}^1 N^T N d\xi d\eta d\zeta \quad (23)$$

$$K = \int_{-1}^1 \int_{-1}^1 \int_{-1}^1 B^T D B |J| d\xi d\eta d\zeta \quad (24)$$

where $|J|$ is the determinant of the Jacobian matrix for the 3-D finite element and ξ , η , and ζ represent coordinates in iso-parametric space.

Fourier superposition. Fourier superposition is a three-step solution process for linear systems that involves a forward transformation into a wavenumber domain, the calculation of a solution to Equation 21 at a number of increments, and the determination of the total solution through an inverse transformation of all incremental solutions. A time-temporal frequency transform pair of a load function p are:

$$p(\omega) = \int_{-\infty}^{\infty} p(t) e^{-i\omega t} dt \quad (25)$$

$$p(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} p(\omega) e^{i\omega t} d\omega \quad (26)$$

Similarly, the distance-spatial frequency (wavenumber) transform pair for expansion in the y-direction are:

$$p(m) = \int_{-\infty}^{\infty} p(y) e^{imy} dy \quad (27)$$

$$p(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} p(m) e^{-imy} dm \quad (28)$$

where

m = wavenumber (spatial circular frequency) in y-direction

Fourier superposition applied in both the time and y-spatial domains leads to:

$$p(m, \omega) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(y, t) e^{-i(\omega t - my)} dt dy \quad (29)$$

$$p(y, t) = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(m, \omega) e^{i(\omega t - my)} d\omega dm \quad (30)$$

The corresponding transformation equation for displacements is:

$$u(y, t) = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u(m, \omega) e^{i(\omega t - my)} d\omega dm \quad (31)$$

Making the load vector specific to steady-state vibrations with constant amplitude, the time-temporal frequency transform pair reduce to:

$$p(\omega) = \bar{p} \quad (32)$$

$$p(t) = \bar{p} e^{i\omega t} \quad (33)$$

where \bar{p} is used to represent amplitude which allows Equations 29 and 30 to be reduced to:

$$p(m)_\omega = \bar{p} \int_{-\infty}^{\infty} p(y) e^{i\omega y} dy \quad (34)$$

$$p(y, t) = \bar{p} \frac{e^{i\omega t}}{2\pi} \int_{-\infty}^{\infty} p(m)_\omega e^{-i\omega y} dm \quad (35)$$

for a specific ω . The corresponding equations for displacements are:

$$u(m)_\omega = \bar{u} \int_{-\infty}^{\infty} u(y) e^{i\omega y} dy \quad (36)$$

$$u(y, t) = \bar{u} \frac{e^{i\omega t}}{2\pi} \int_{-\infty}^{\infty} u(m)_\omega e^{-i\omega y} dm \quad (37)$$

Element condensation. The process of element condensation is the key aspect of the reduction of computational effort. Element condensation refers to the process of reducing the number of degrees of freedom by relating points adjacent in the y-direction using the functional relationship of the Fourier expansion. The dependent degrees of freedom are then eliminated by expressing them in terms of the degrees of freedom of the in-plane nodes. In this case, the degrees of freedom corresponding to the nodes outside of the x-z plane are eliminated. Each node in the 2-D mesh maintains three degrees-of-freedom.

Consider an arbitrary discretized model of a physical system that meets the requirement of uniform geometry and material properties in one direction. The coordinate system is chosen to have the z-direction positive down and the other in-plane direction to be x. Consider three vertical planes separated by a distance of Δy at some arbitrary location along the geosystem. The 3-D dynamic stiffness matrix for any element between the slices, such as that shown in Figure 1a, is calculated using Equations 23, 24, and then 22.

The dynamic stiffness matrix for a single element, such as that shown in Figure 1a, can be partitioned as:

$$\bar{S} = \begin{bmatrix} \bar{S}_{11} & \bar{S}_{12} \\ \bar{S}_{21} & \bar{S}_{22} \end{bmatrix} \quad (38)$$

where the subscripts "1" and "2" refer to the degrees of freedom on the positive and negative face in the y-direction, respectively. The assemblage of the dynamic equations for any two finite elements adjacent in the y-direction, as shown in Figure 1b, can be reduced by canceling the time-dependent exponential term on each side to:

$$\begin{bmatrix} \bar{S}_{11}^+ & \bar{S}_{12}^+ & 0 \\ \bar{S}_{21}^+ & \bar{S}_{22}^+ + \bar{S}_{11}^- & \bar{S}_{12}^- \\ 0 & \bar{S}_{21}^- & \bar{S}_{22}^- \end{bmatrix} \begin{Bmatrix} \bar{U}_b \\ \bar{U}_a \\ \bar{U}_c \end{Bmatrix} = \begin{Bmatrix} \bar{P}_b \\ \bar{P}_a \\ \bar{P}_c \end{Bmatrix} \quad (39)$$

where

- "+" denotes element in positive y-direction (from Δy to 0)
- "-" denotes element in negative y-direction (from 0 to $-\Delta y$)
- "a" denotes the degrees of freedom on face a (i.e., at $y = 0$)
- "b" denotes the degrees of freedom on face b (i.e., at $y = +\Delta y$)
- "c" denotes the degrees of freedom on face c (i.e., at $y = -\Delta y$)

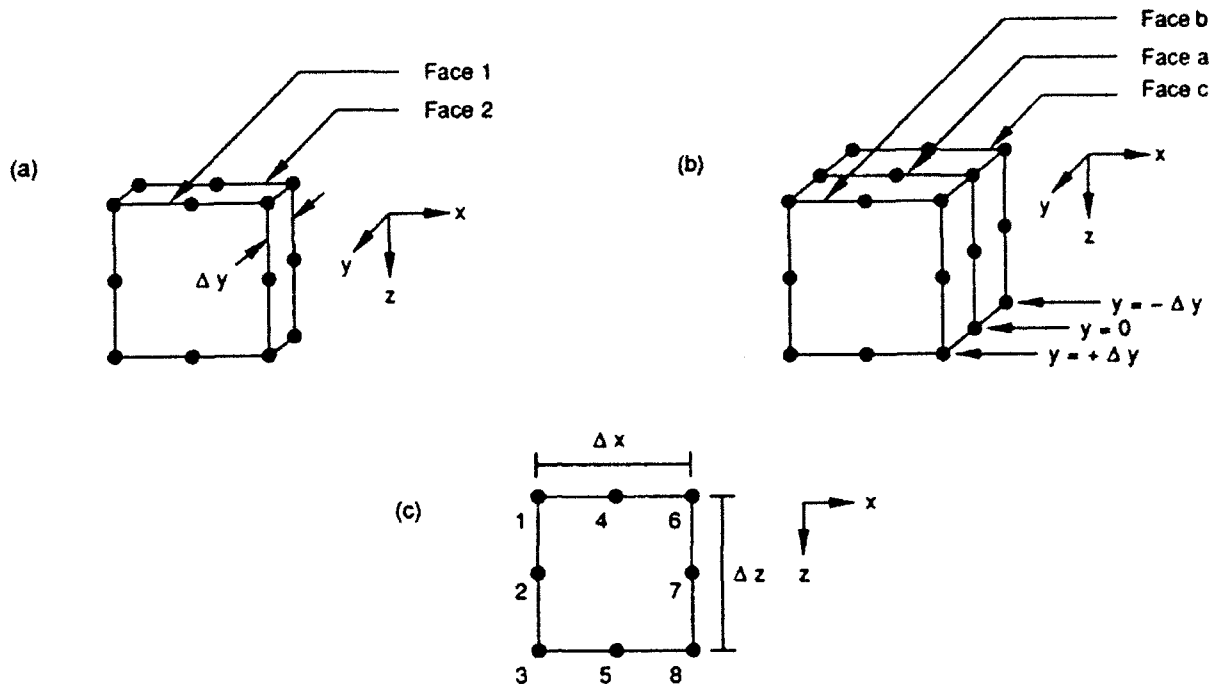


Figure 1. Condensation of finite elements adjacent in out-of-plane (y) direction

Using the Fourier expansion described earlier (Equations 34 and 36), forces and displacements are expressed as:

$$\hat{\mathbf{P}}(m) = \int_{-\infty}^{\infty} \bar{\mathbf{P}}(y) e^{imy} dy \quad (40)$$

$$\hat{\mathbf{U}}(m) = \int_{-\infty}^{\infty} \bar{\mathbf{U}}(y) e^{imy} dy \quad (41)$$

where $\hat{\mathbf{P}}$ and $\hat{\mathbf{U}}$ are used to represent vectors of nodal forces and displacements, respectively, in m space. Rewriting Equation 39 to incorporate the Fourier expansion of loads:

$$\begin{bmatrix} \bar{s}_{11}^+ & \bar{s}_{12}^+ & 0 \\ \bar{s}_{21}^+ & \bar{s}_{22}^+ + \bar{s}_{11}^- & \bar{s}_{12}^- \\ 0 & \bar{s}_{21}^- & \bar{s}_{22}^- \end{bmatrix} \begin{Bmatrix} \hat{\mathbf{U}}_b(m) \\ \hat{\mathbf{U}}_a(m) \\ \hat{\mathbf{U}}_c(m) \end{Bmatrix} = \begin{Bmatrix} \hat{\mathbf{P}}_b(m) \\ \hat{\mathbf{P}}_a(m) \\ \hat{\mathbf{P}}_c(m) \end{Bmatrix} \quad (42)$$

In the transform (m) space, the displacements on the "b" and "c" faces are related to the displacements on the "a" face at any instant in time by the simple relationships:

$$\hat{\mathbf{U}}_b(m) = \bar{\mathbf{U}}_a(m) e^{-i\Delta y} \quad (43)$$

$$\hat{U}_n(m) = \bar{U}_n(m) e^{-im\Delta y} \quad (44)$$

Defining:

$$\hat{S}(m) = \bar{S}_{21} e^{-im\Delta y} + (\bar{S}_{11} + \bar{S}_{22}) + \bar{S}_{12} e^{im\Delta y} \quad (45)$$

Equations 43 and 44 can be substituted into Equation 42 to get the system of equations for the equivalent two-dimensional system shown in Figure 1c:

$$\hat{S}(m) \hat{U}_n(m) = \hat{P}_n(m) \quad (46)$$

This formulation, then, allows the 3-D finite element with a 2-D geometry to be represented with an equivalent 2-D finite element. The representation of surface loads are described below.

Surface loads. This study focuses on the preparation for analysis of waves propagating from a synthetic, 3-D source. Vibroseis trucks generally use a rectangular platen with plan dimensions on the order of 1 by 2 m (3 by 7 ft). At large distances from the source and with large wavelengths, this area approaches a point source. Therefore, the horizontal distributions of the load considered for this study were a point load and a rectangular load of various sizes. A point source is not a physical reality and is difficult to replicate with finite elements. Kang (1990) used a point load and circular load as these were appropriate vibration sources for pavement systems.

The formulation for equivalent nodal forces in the x-direction for point and rectangular loads are described below. The formulation of equivalent nodal forces for rectangular loads involves integration of the force distribution in light of the interpolation function:

$$\bar{P} = \int_x N^T \bar{p} dx \quad (47)$$

The distribution of forces applied to the platen is assumed to be uniform and therefore the integration reduces to simple algebra. For example, a continuous, uniform load with a total magnitude of unity ($p \cdot \Delta x = 1$), the equivalent nodal forces are 1/6 for the endpoints and 4/6 for the midpoint.

Time-dependent displacements. The real-valued, time-dependent displacements may be obtained from the calculated complex displacements, U . If the forcing function is of the form $\sin \omega t$, then:

$$u_i = A_i \sin(\omega t) + B_i \cos(\omega t) \quad (48)$$

If the forcing function is of the form $\cos \omega t$, then:

$$u_i = A_i \cos(\omega t) - B_i \sin(\omega t) \quad (49)$$

where

A_i = real part of complex displacement amplitude at node i

B_i = imaginary part of complex displacement amplitude at node i

For the analysis of the vibrations produced by a Vibroseis, Equation 49 is more appropriate. The phase angle of motion, ϕ , is calculated by:

$$\phi = \tan^{-1} \left(\frac{B_1}{A_1} \right) \quad (50)$$

VALIDATION STUDIES

Validation studies and parametric analyses were used to prove that the formulation and computer implementation are sound, accurate, and stable for the limited problem class to which accurate solutions are available. The findings of validation studies are not mutually exclusive from the parametric analyses because the definition of the problems for validation should conform somewhat to the findings of parametric analyses. The results of the validation studies are described below; the parametric analyses are described in the next section.

The best form of validation consists of comparing the results between a subject program and exact mathematical relationships for several different problems. Comparisons with measured data or prototype testing provide a constructive means to confirm findings when conducted under certain controlled conditions. These comparisons are not appropriate as the primary means of validation, however. Comparisons with other numerical approximations are even less appropriate for validation. Validation of the computer code developed for this study was made through comparisons with analytical results for the simplest class of planar geometry -- a horizontally layered system extending to infinity. Green's function solutions formulated for axi-symmetric problems by Kausel (1981) were used exclusively. Some minor differences in displacement may exist between the Green's function solutions and the 2-D approximations because the shape of the load is different -- disk loads were used for the axi-symmetric problem and square loads were used for this study. The same total area and total load of unity were used to minimize these differences. The model systems used to validate the computer code are described in the next section.

The validation studies described in this part pertain to variations in system geometry, material properties, and aspects of finite element analysis. The dynamic vertical displacements are of primary interest because they predominate in surface motions caused by vertical excitations. Moreover, vertical vibrations are normally measured in non-destructive testing techniques such as the SASW method. For purposes of this paper, the results are presented in terms of the variations of real and imaginary components of dynamic displacement in the y-direction (calculated at node beneath the centroid of the load and expanded out in the y-direction). Comparisons are made at the free (ground) surface with distances normalized to the wavelength of Rayleigh waves, λ , for Model 1. The displacements are oriented positive-down to be consistent with the convention used in the formulation and correspond to the top surface ($z = 0$).

Test Models and Discretization Schemes

Four hypothetical models were created for validation studies and are shown with unit-less dimensions in Figure 2. These models were designed to represent ideal site conditions of horizontally layered soil overlying rock and realistic material properties (considering units of ft-lb-sec) while

conforming to limitations of the analytical solutions. All models have the same total height (1000 units) and are assumed to overlay a rigid material. Model 1 is the simplest system -- a homogeneous medium overlying rock. The range of material properties for this medium used in the following comparisons are shown in Figure 2. The other three models consist of four homogeneous layers overlying rock with different combinations of stiffness.

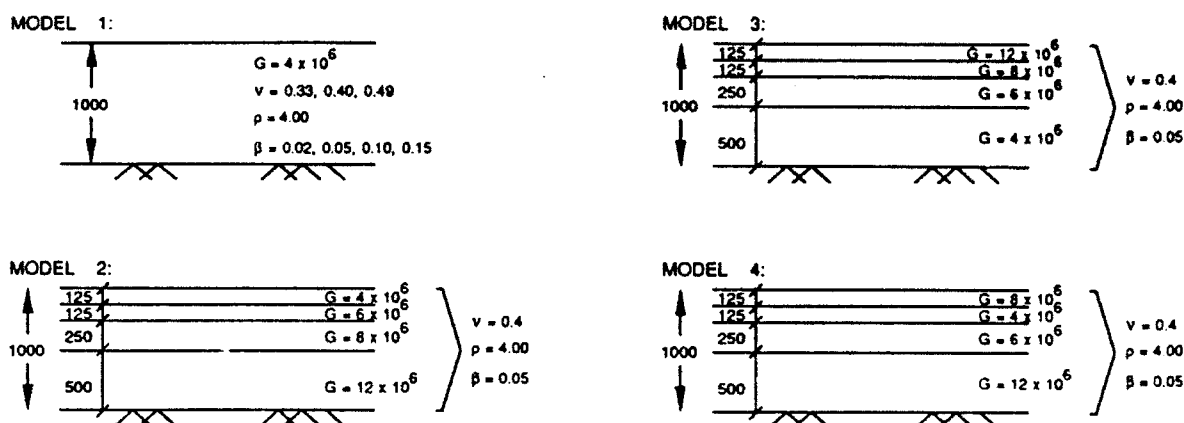


Figure 2. Test models used for validation studies

A domain with dimensions 1000 units high and 2500 units wide was chosen, along with the material properties and frequency of excitation, to be large enough to ignore the effects of reflections and correspond to about 3λ high by 8λ wide. Three different finite element meshes were created to represent this domain, using 4 by 10, 8 by 20, and 16 by 40 square elements. The size of these elements corresponds to 0.8λ , 0.4λ , and 0.2λ , respectively. A plane of symmetry at the left boundary, defined by $x = 0$, was utilized to reduce the degrees of freedom by nearly one-half. A frequency of excitation of 3 Hz, system damping of 2 percent, and a radius of load of 5.64 (total area of 100) were used to analyze all four models.

Analytical Solutions

The Green's function solutions formulated by Kausel (1981) were calculated with the computer code *PUNCH* (Kausel 1989) using a personal computer. The calculated solution approaches the exact solution as the number of layers increases. Twenty-five was found to be an adequate number of layers to adequately represent these models for further comparisons and validation. The displacements calculated using *PUNCH* correspond to a disk load with radius r and total load, P , of 1 ($= \pi r^2$) or a point load with magnitude of unity.

Element Performance to Static Loads

The specialized 3-D finite element was evaluated for the ability to represent static response to various loads. This evaluation was accomplished by comparing the results of two approaches with analytical solutions. One approach was to place the algorithms defining the element stiffness into a static finite element computer code and examine the response of a cantilever beam. The other approach used *vib3* with a point load acting on a homogeneous body with the frequency equal to 0. Each of these are described below.

Static finite element code. A static finite element code was used to evaluate the specialized finite element. This program evolved from an unnamed finite element code used for instructional purposes at the University of Texas at Austin. A cantilever beam was discretized with 2, 5, 10, 40, and 80 elements and subjected to tension, compression, and shear-induced bending loads. The effect of element shape was also evaluated by considering square, rectangular, parallelogram, and trapezoidal configurations. Comparisons between calculated and closed-form solutions for displacements and stresses were good and indicate that the algorithms defining the element stiffness are accurate for conditions of static loading.

Dynamic code. The static vertical displacements calculated using vib3 with Model 1 at the ground surface is shown in Figure 3 using a mesh that was 16 elements high by 40 elements wide. The comparisons with Green's function solutions are excellent for the real part and very good for the imaginary part at distances slightly removed from the point of load (greater than 100 units). Comparisons are similar at all depths. The imaginary part should be zero at all distances but vib3 produces non-zero values at locations close to the load. The less favorable comparisons near the point of loading are common when modeling a point load using the finite element method. These errors are normally minimized through mesh refinement near the point of loading but accuracy close to the source is not of interest for this study.

Approximations for Dynamic Loads

The computer code vib3 was used to calculate dynamic displacements for each of the four models described previously. These results were then compared with the Green's function solutions presented in the previous section. All four models were discretized using the finest mesh. A square load with plan dimensions of 5 by 5 centered about the origin with a total load of 1 was applied at a frequency of 3 Hz. The wavelength for Rayleigh waves is then about 31λ and the dimension of the square elements are 62.5 units or about 0.2λ . The material properties are listed in Figure 1.

The parameters defining the condensation and Fourier expansion for the validation were selected based on the findings of Kang (1990). Values of $\Delta y = 0.05\lambda$ and the number of Fourier discretization points, NM, equal to 256 were fixed for the comparisons and the finest finite element mesh was used unless otherwise specified. This provided for a discretized extent (in the y-direction) of -13λ ($\pm 6.4\lambda$), slightly less than the total extent discretized in the x-direction ($\pm 8\lambda$). Displacements at distances up to 5λ , or about 1500 units, are used for comparison because the amplitudes are rather small beyond this distance.

Model 1: Homogeneous system. The results for Model 1 at $z = 0$ is shown in Figure 4 and compared with the Green's function solutions. The variation of the real part of the complex displacements compare well with the Green's function solution. The variation of the imaginary part closely follows the Green's function solutions. Both parts of the calculated solutions compare more favorably at distances less than 3λ , about half the distance expanded in the y-direction. The comparisons are also slightly better at some depth as compared to at the ground surface.

The effects of varying Poisson's ratio and damping ratio on the displacements for Model 1 were also examined. Comparisons between the calculated and the Green's function solutions for variations in ν indicates that the 2-D approximation provides a reasonable means of representing different ν . The best comparison is for $\nu = 0.40$ and the poorest comparison

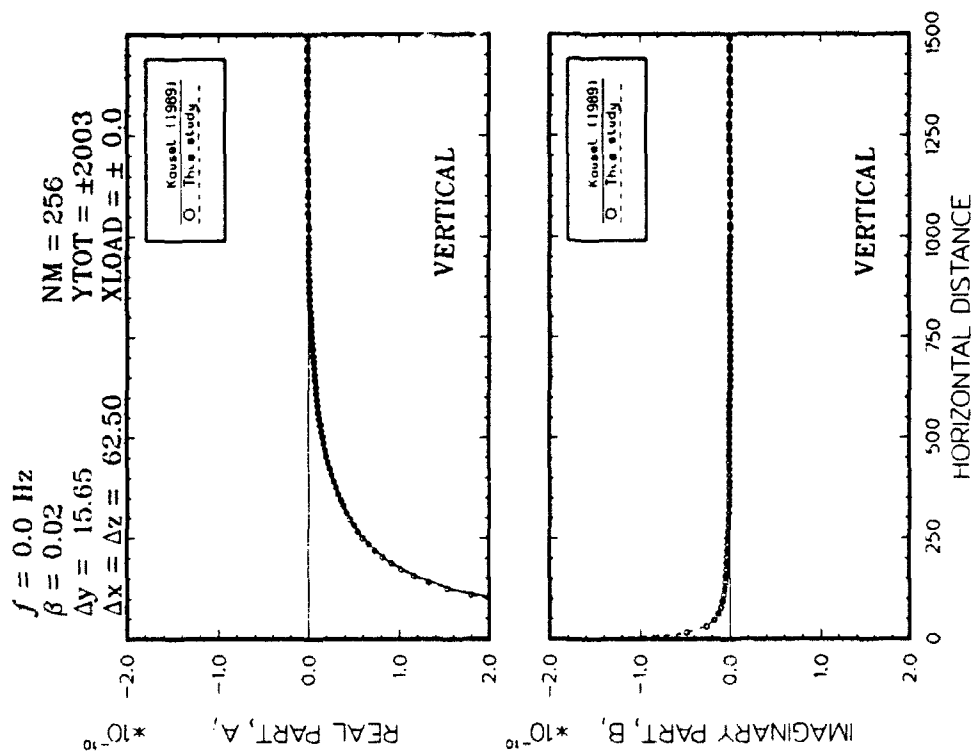


Figure 3. Vertical displacements at $z = 0$ for static point load and comparison with Green's function solution

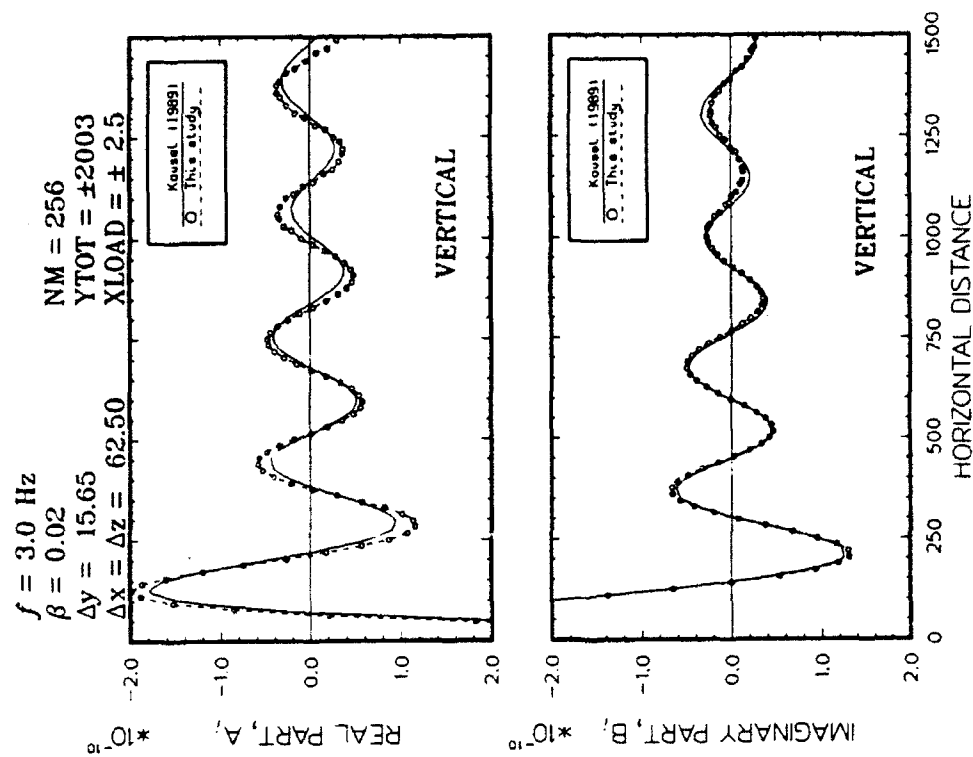


Figure 4. Vertical displacements for Model 1 with Green's function solution

is for $\nu = 0.49$. Generally, the imaginary part compares well but the real part varies somewhat. The 2-D approximation provides an accurate means of representing damping. The accuracy of calculated displacements improves somewhat as the damping ratio increases. The results for 5 percent damping compare much better with the Green's function solution than the results for 2 percent. Both 2 and 5 percent damping levels are used for comparisons hereafter.

Other models: Stiffness varying with depth. The results for Models 2, 3 and 4 are shown in Figures 5 through 7 and compared with the Green's function solutions. Total distances are used rather than normalized distances since considerable dispersion is expected. The results for vertical displacements are nearly equivalent to the Green's function solutions except at the first peak in the real part for all three cases.

PARAMETRIC ANALYSES

Parametric analyses were conducted to assess the sensitivity of the formulation and computer code *vib3* to anticipated ranges of system variables. Calculations were made using Model 1 and the finest mesh except in the case of examining sensitivity to mesh size. Green's function solutions calculated using *PUNCH* (Kausel 1989) are used for comparison.

Effect of Δy

The effect of the spatial increment of discretization in the y-direction was evaluated by comparing the results using three values of Δy between 0.05λ and 0.20λ (0.05λ used for validation study). The number of FFT points, NM, was also varied to keep the total discretized distance in the y-direction, YTOT, constant. This distance is defined by:

$$YTOT = NM \cdot \Delta y \quad (51)$$

Keeping YTOT constant serves to isolate the effects of Δy . The variation of vertical displacements for the different values of Δy are compared in Figure 8.

The large difference among relationships presented in Figure 8 indicate that Δy has a significant effect on the ability of *vib3* to accurately calculate dynamic displacements. Comparisons between the calculated displacements and Green's function solutions are favorable when $\Delta y \leq 0.10\lambda$ although some improvement is noticeable by decreasing Δy to 0.05λ . These results are consistent with Kang's (1990) who recommended that $\Delta y \leq 0.10\lambda$. The results for $\Delta y \geq 0.20\lambda$ are considered to be too inaccurate. The parametric analysis of Δy with respect to λ indirectly addresses the effect of frequency of excitation on the results. For a homogeneous system (with constant stiffness), λ is inversely proportional to frequency. So, the spatial increment Δy can also be put in terms of frequency:

$$\Delta y \leq \frac{v}{10 f} \quad (52)$$

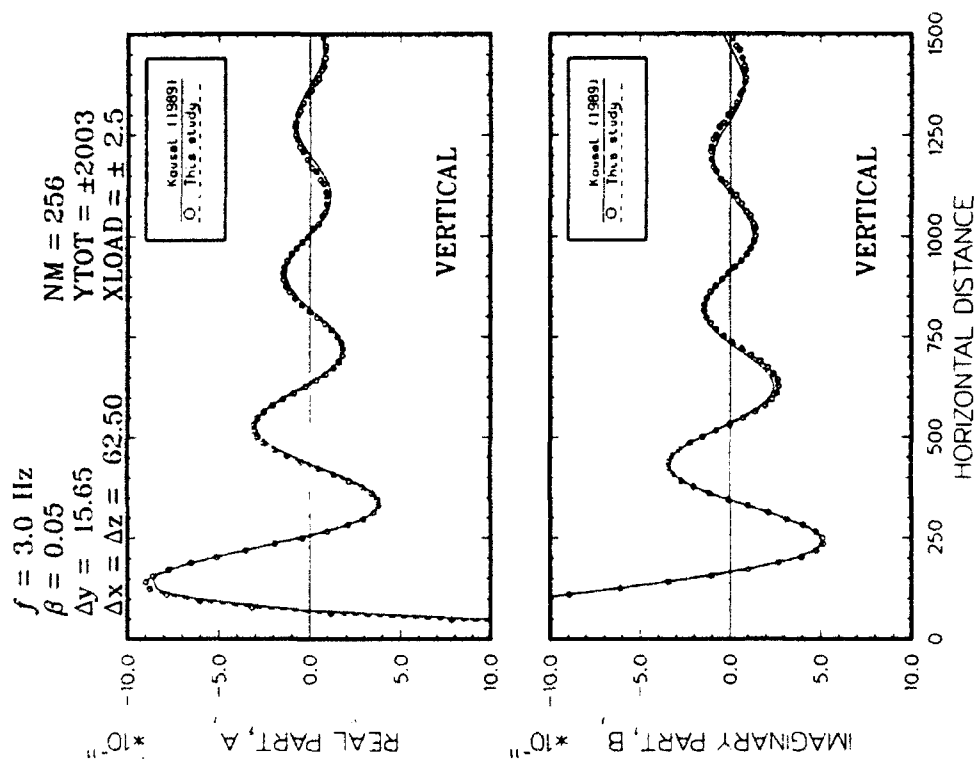


Figure 7. Vertical displacements for Model 4 with Green's function solution

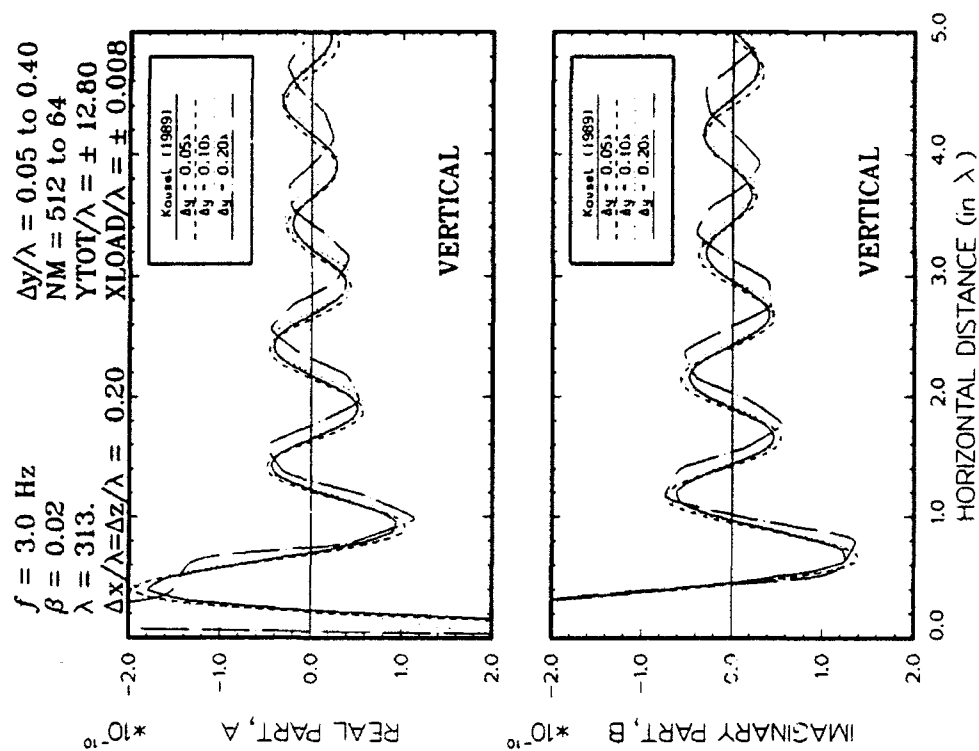


Figure 8. Comparison of vertical displacements showing effect of Δy with Green's function solution

where

V = phase velocity

f = frequency (Hz)

The phase velocity can be taken equal to the Rayleigh wave velocity as a first approximation. Similar relationships to Equation 52 have been observed in other types of discretized solutions for dynamic loading.

Effect of Extent of Fourier Expansion

The comparisons for the effect of Δy were made using a constant value of YTOT. The effect of varying YTOT was examined next. The total distance was varied at three values between 3.2λ and 12.8λ (corresponding to $\pm 1.6\lambda$ and $\pm 6.4\lambda$, respectively) by keeping Δy constant at 0.05λ and varying NM between 128 and 512. The variation of vertical displacements for the three values of YTOT are compared in Figure 9. The results for the vertical displacements are very good for the case of $YTOT = \pm 12.8\lambda$. The results for $YTOT = \pm 6.4\lambda$ are also good, especially for distances less than 3λ , and the results for $YTOT = \pm 3.2\lambda$ are considered to be too inaccurate. A threshold of 10λ is likely to be appropriate.

Effect of Element Size

The effect of varying the size of the finite elements on the solution determined using the three different meshes is shown in Figure 10. The values of Δx ($= \Delta z$) corresponding to these three meshes are 0.20λ , 0.40λ , and 0.80λ . The results for the variation for the three meshes are compared in Figure 10.

The variation of vertical displacements compare well with the Green's function solutions except for the coarsest mesh (4 by 10 elements). The results for the coarsest mesh are unacceptable. The finest mesh produces peak values of displacement slightly greater than the Green's function solution and the original mesh.

Effect of Width of Load

The load width in the x- and y-directions, XLOAD, ranges from a point load to $\pm 0.064\lambda$ (80 by 80 in total plan dimensions at 3 Hz). For all practical purposes and at these distances and depths, these loads are essentially point loads. The results for the variation for the different load widths are compared in Figure 11. Little noticeable effect is evident as XLOAD is varied over the specified range. A threshold of load $\leq \pm 0.10\lambda$ appears to be reasonable to maintain good accuracy. The small difference between the results for the point load and the smallest square load is somewhat surprising. Kang (1990) noticed a larger difference and researchers have recognized the difficulty in calculating an accurate distribution of displacements from a point load using the finite element procedure without a refined mesh in the vicinity of the load.

Computational Effort

The amount of time necessary to run the program with different system parameters was reviewed. The two parameters considered to have the greatest effect are the number of FFT points, NM, and the number of degrees of freedom, dof. (Recall that the equations are solved for only half of the NM and the results mirrored prior to the inverse Fourier transform.) Comparisons of user CPU (central processing unit) times versus NM are shown in Figure 12. The three finite element meshes described earlier were used to provide a range in

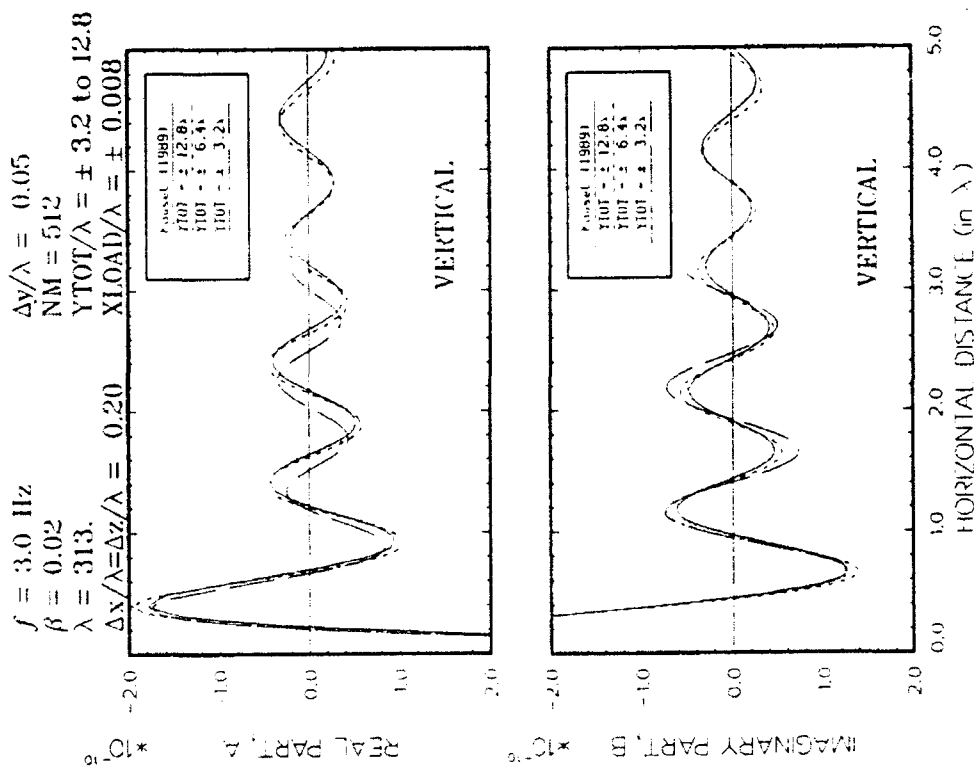


Figure 9. Comparison of vertical displacements showing effect of YTOT with Green's function solution

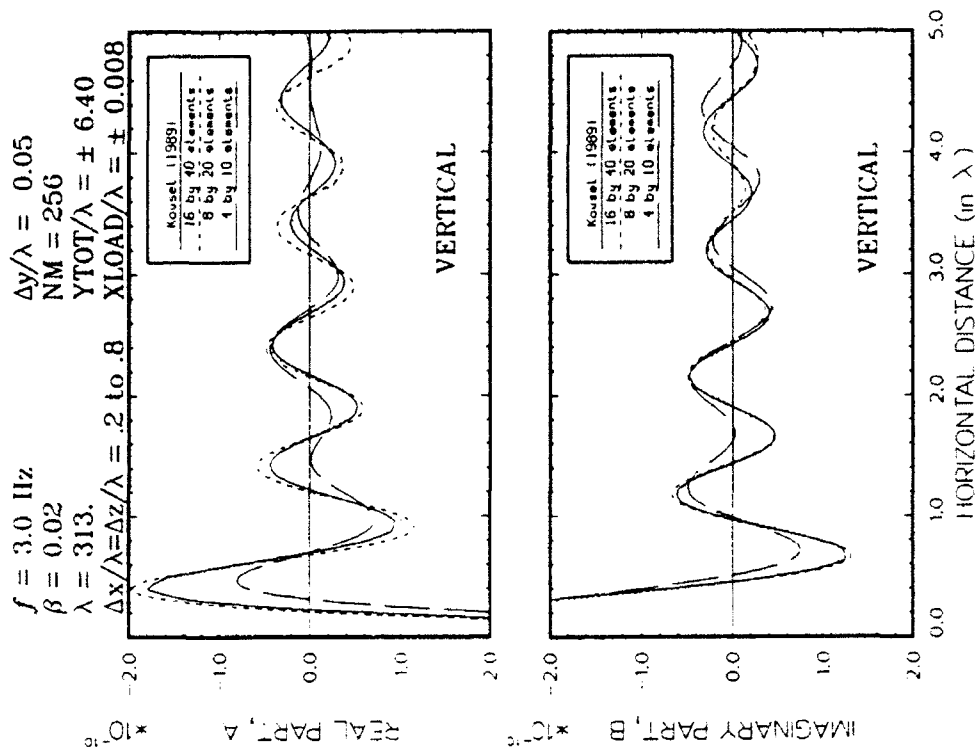


Figure 10. Comparison of vertical displacements showing effect of size of finite elements with Green's function solution

degrees of freedom. The solution times are of the same order as NM (linear relationship) for a fixed number of dof. The slopes of these lines range from 1.3 to 26. Comparisons of user CPU times versus dof for various NM are shown in Figure 13. The relationship is slightly non-linear for a fixed NM; the exponent of dof is about 1.12 and increases slightly as NM increases.

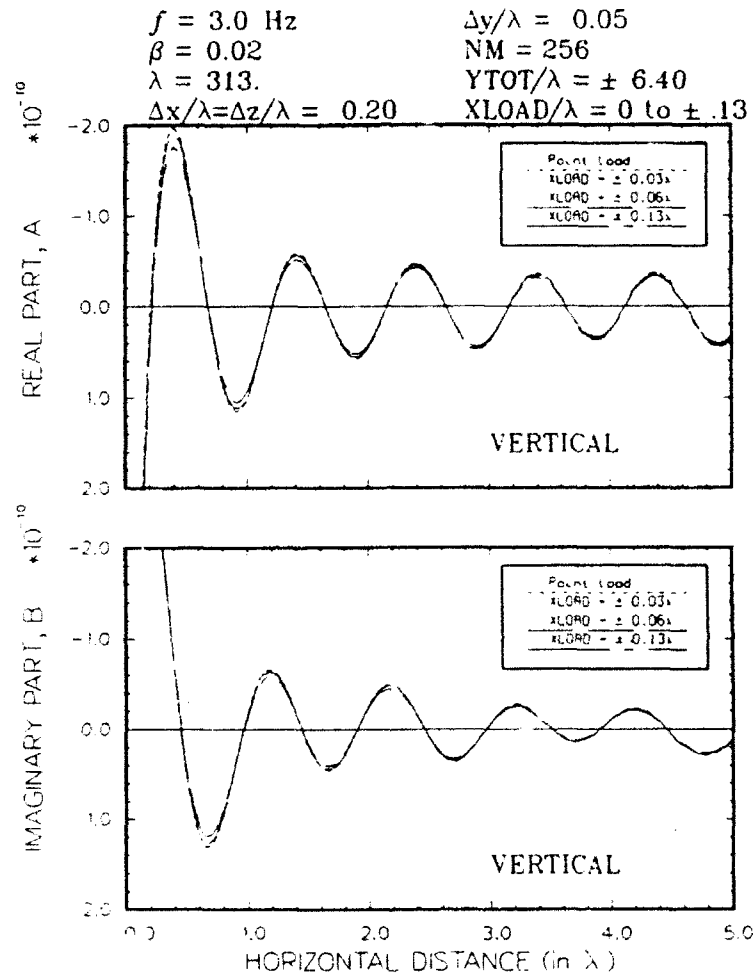


Figure 11. Comparison of vertical displacements showing effect of varying load width

The amount of time saved in using the present formulation over a conventional 3-D finite element formulation was estimated by solving the problem for Model 1 using the commercial software package ABAQUS. Two planes of symmetry were used such that only a 3-D quarter space was required to be discretized. A total discretized space of 8λ by 3λ in plan by 3λ deep was used and the element size was equal to that used in the 8 by 20 mesh ($\Delta x = \Delta y = \Delta z = 0.40\lambda$). A 3-D isoparametric element with 20 nodes (quadratic interpolation functions in all three directions) was selected as being comparable. The extent and accuracy of discretization in the y-direction is roughly equivalent to $NM = 64$ and $\Delta y = 0.10\lambda$ which were used with vib3. Free end conditions were used for non-symmetric boundaries. The calculated results

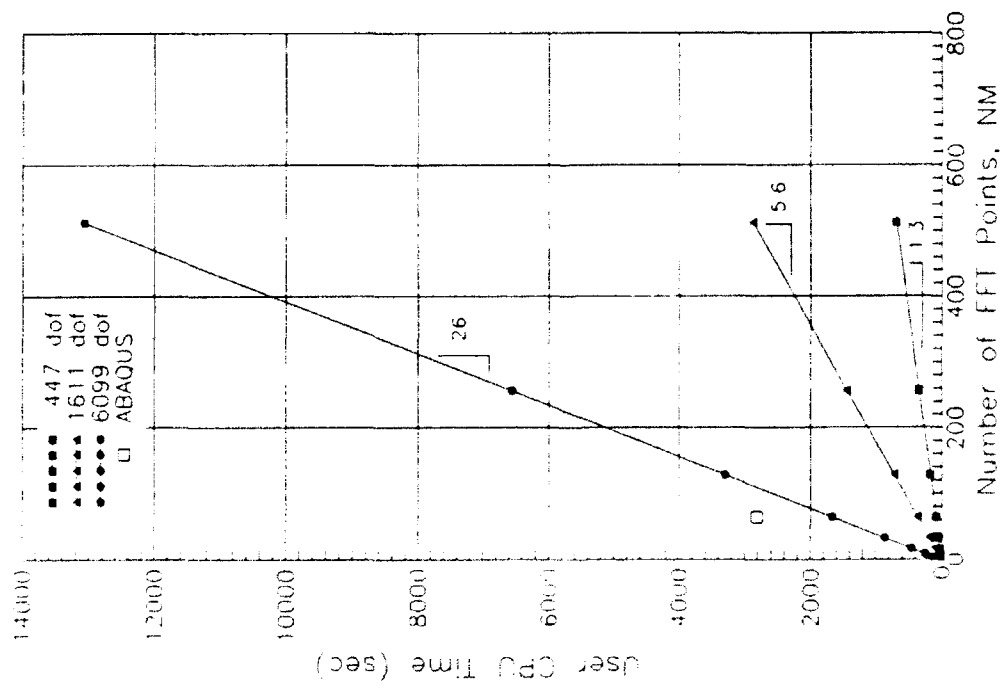


Figure 12. Comparison of CPU execution times on CRAY Y-MP versus number of FFT points

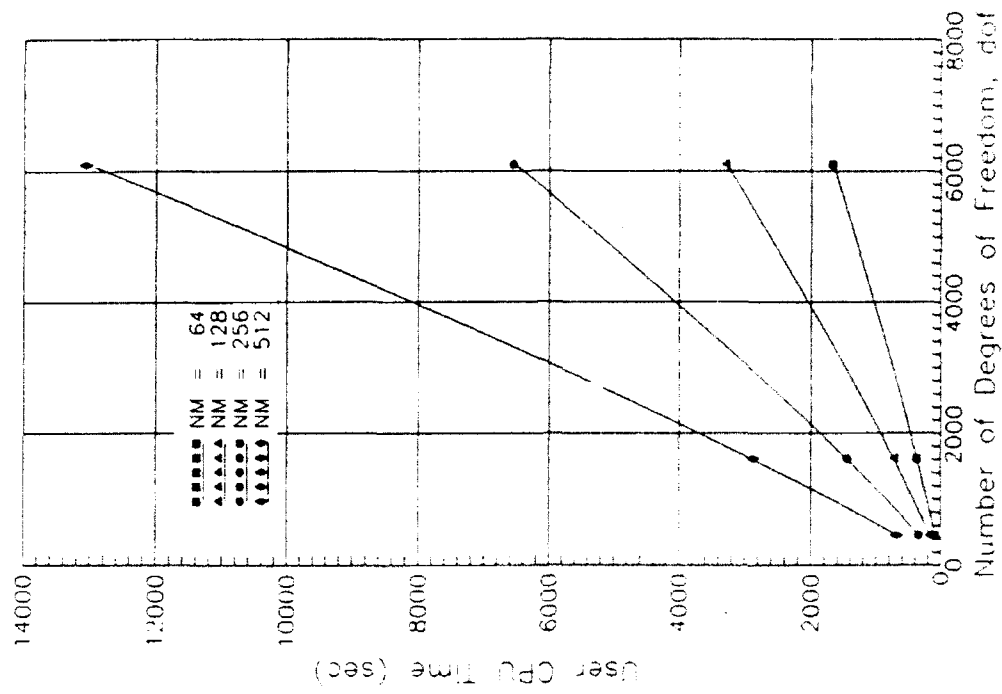


Figure 13. Comparison of CPU execution times on CRAY Y-MP versus number of degrees of freedom

were not of particular interest. The user CPU time required by ABAQUS to solve for dynamic displacements was about 2820 sec compared to 370 sec using vib3. A comparison of times is shown in Figure 12. Moreover, almost 8 Mwords of memory were required to solve the problem using ABAQUS whereas about 3.6 Mwords were used by vib3.

SUMMARY AND CONCLUSIONS

A method to calculate dynamic displacements in 2-D geosystems produced by a harmonic point or rectangular load has been formulated and implemented in a 2-D finite element computer code that functions on the US Army CRAY supercomputer at WES. The formulation involves creating a 3-D dynamic stiffness matrix and then condensing the components into an equivalent 2-D dynamic stiffness matrix. The out-of-plane loads are represented by a Fourier expansion and applied as nodal forces. The solution to the system of equations is made for each spatial wavenumber and then the inverse Fourier transform produces the dynamic displacements. The computer code has been validated with analytical solutions for the case of axi-symmetric geosystems subjected to static and dynamic loads. Parametric studies were performed to determine how the accuracy of the calculated displacements are affected by the various input parameters. All comparisons indicate that this method is a viable alternative to more time consuming 3-D numerical solution methods.

Validation studies were performed for cases of static and dynamic loads generally using reasonable values of system parameters. The effects of static loads were examined in terms of displacement and stress field for cantilever beams in tension, compression, and torsion using the specialized 16-node, 3-D, finite element incorporated into a static 3-D finite element computer code. Calculated values were compared with closed-form elastic solutions. The displacements produced by static and dynamic point and square loads were examined for cases of a homogeneous medium and three combinations of four-layered media using vib3 and compared with Green's function solutions proposed by Kausel (1981).

The analysis of parameters necessary to the program indicates that once threshold values are met, the formulation is stable to variations in parameters defining the discretization, condensation, and Fourier expansion of the problem. These thresholds are: $\Delta y \leq 0.05\lambda$, $YTOT \geq \pm 10\lambda$, $\Delta x = \Delta z \leq 0.30\lambda$ (for quadratic interpolation), and $XLOAD \leq \pm 0.10\lambda$. Additional improvements may be realized by using even smaller values of Δx , Δy , and Δz . Displacements can be calculated about 8 times faster using the new formulation when compared to the 3-D finite element code ABAQUS.

REFERENCES

- Hanazato, T., Ugai, K., Mori, M., and Sakaguchi, R. 1991. "Three-Dimensional Analysis of Traffic-Induced Ground Vibrations, J. Geotech. Engr., Vol 117, No. 8, pp 1133-1151.
- Hardin, B. O. and Drnevich, V. P. 1972. "Shear Modulus and Damping in Soils: Measurement and Parameter Effects," J. Geotech. Engr., Vol 98, No. 6, pp 603-624.
- Johnston, D. H., Toksöz, M. N., and Timur, A. 1979. "Attenuation of Seismic Waves in Dry and Saturated Rocks: II. Mechanisms," Geophysics, Vol 44, pp 691-711.
- Kang, Y. V. 1990. "Effect of Finite Width on Dynamic Deflections of Pavements," PhD dissertation, Univ. Texas, Austin, TX.

- Kausel, E. 1981. "Explicit Solution for the Green Functions for Dynamic Loads in Layered Media," MIT Research Report R81-13, Cambridge, MA.
- Kausel, E. 1989. "PUNCH: Program for the Dynamic Analysis of Layered Soils," ver. 3.0, Massachusetts Institute of Technology, Cambridge, MA.
- Lai, J. Y. and Booker, J. R. 1991. "Application of Discrete Fourier Series to the Finite Element Stress Analysis of Axi-Symmetric Solids," Int'l J., Num. Meth. Engr., Vol 31, pp 619-647.
- Lin, H-T. and Tassoulas, J. L. 1987. "Discrete Green Functions for Layered Strata," Int'l J. Num. Meth. Engr., Vol 24, pp 1645-1658.
- Love, A. E. H. 1944. A Treatise on the Mathematical Theory of Elasticity, 4th ed., Dover, New York, NY.
- Nazarian, S. and Stokoe, K. H. II, 1985a. "In Situ Determination of Elastic Moduli of Pavement Systems by Spectral-Analysis-of-Surface-Waves Method (Practical Aspects)," Rpt. 368-1F, Cntr. Trans. Res., Austin, TX.
- Nazarian, S. and Stokoe, K. H. II, 1985b. "In Situ Determination of Elastic Moduli of Pavement Systems by Spectral-Analysis-of-Surface-Waves Method (Theoretical Aspects)," Rpt. 437-1, Cntr. Trans. Res., Austin, TX.
- Rix, G. J. and Leipski, E. A. 1991. "Accuracy and Resolution of Surface Wave Inversion," Recent Advances in Instrumentation, Data Acquisition, and Testing in Soil Dynamics, ASCE, Geot. Pub. 29, ed. Bhatia & Blaney, pp 17-32.
- Runesson, K. and Booker, J. R. 1982. "Efficient Finite Element Analysis of 3D Consolidation," Numerical Methods Geomechanics, Edmonton 1982, ed. Eisenstein, Balkema Press, Rotterdam, pp 359-364.
- Runesson, K. and Booker, J. R. 1983. "Finite Element Analysis of Elastic-Plastic Layered Soil Using Discrete Fourier Series Expansion," Int'l J. Num. Meth. Engr., Vol 19, pp 473-478.
- Sykora, D. and Roesset, J. 1992. "Two-Dimensional Planar Geosystems Subjected to Three-Dimensional Dynamic Loads," Technical Rpt. GL-92-16, US Army Engineer Waterways Experiment Station, Vicksburg, MS.
- Timoshenko, S. P. and Goodier, J. N. 1970. Theory of Elasticity, McGraw-Hill, New York, NY.
- Toksöz, M. N., Johnston, D. H., and Timur, A. 1979. "Attenuation of Seismic Waves in Dry and Saturated Rocks: II. Laboratory Measurements," Geophysics, Vol 44, pp 681-690.
- Winnicki, L. A. and Zienkiewicz, O. C. 1979. "Plastic (or Visco-Plastic) Behaviour of Axisymmetric Bodies Subjected to Non-Symmetric Loading -- Semi-Analytical Finite Element Solutions," Int'l J., Num. Meth. Engr., Vol 14, pp 1399-1412.
- Wolf, J. P. 1985. Dynamic Soil-Structure Interaction, Prentice-Hall, Inc., Englewood Cliffs, NJ, pp 15-16.
- Zienkiewicz, O. C. and Taylor, R. L. 1989. The Finite Element Method, Vol 1: Basic Formulation and Linear Problems, McGraw-Hill, London, 4th ed.

ACKNOWLEDGEMENTS

The analyses described herein, unless otherwise noted, were conducted under the In-House Laboratory Independent Resesearch (ILIR) Program of the United States Army Engineers by the Waterways Experiment Station. The views of the authors do not purport to reflect the position of the Department of the Army or the Department of Defense. Permission was granted by the Chief of Engineers to publish this information.

ABAQUS is a registered trademark of Hibbit, Karlsson and Sorensen, Inc., Providence, RI.

GENERALIZED STROH FORMALISM FOR ANISOTROPIC ELASTICITY FOR GENERAL BOUNDARY CONDITIONS*

T. C. T. Ting and M. Z. Wang
Department of civil Engineering, Mechanics and Metallurgy
University of Illinois at Chicago
Box 4348, Chicago, IL 60680

ABSTRACT. The Stroh formalism for two-dimensional deformations of anisotropic elastic bodies is generalized so that it can be easily applied to boundary conditions of the type more general than the prescription of traction, displacement or slip boundary conditions. A simple modification is all it takes to encompass all eight different types of boundary conditions. The final solution to a given problem, which looks very similar to that of unmodified version, is applicable to anyone of the eight boundary conditions. By relaxing the definition of I_u and I_ϕ defined in the paper, the number of different types of boundary conditions is increased to cover more than eight types. It is worth mentioned that the modifications required on the Stroh formalism are very minor. Yet the results are applicable to a rather wide range of boundary conditions.

EXTENDED SUMMARY. The present work is motivated by, and is an extension of, the work presented in [1] in which the boundary conditions at $x_2 = \pm 1$ of a semi-infinite strip of anisotropic elastic material can be anyone of the following eight conditions:

$$\sigma_{21} = 0, \quad \sigma_{22} = 0, \quad \sigma_{23} = 0. \quad (1a)$$

$$\sigma_{21} = 0, \quad \sigma_{22} = 0, \quad u_3 = 0. \quad (1b)$$

$$\sigma_{21} = 0, \quad u_2 = 0, \quad \sigma_{23} = 0. \quad (1c)$$

$$\sigma_{21} = 0, \quad u_2 = 0, \quad u_3 = 0. \quad (1d)$$

$$u_1 = 0, \quad \sigma_{22} = 0, \quad \sigma_{23} = 0. \quad (1e)$$

$$u_1 = 0, \quad \sigma_{22} = 0, \quad u_3 = 0. \quad (1f)$$

$$u_1 = 0, \quad u_2 = 0, \quad \sigma_{23} = 0. \quad (1g)$$

$$u_1 = 0, \quad u_2 = 0, \quad u_3 = 0. \quad (1h)$$

*Supported by the U. S. Army Research Office.

Equations (1a), (1h) and (1c) are, respectively, the conditions for a traction-free, rigid, and slip boundary. According to the Stroh formalism [2, 3] the stresses σ_{ij} are related to the stress function ϕ_i by

$$\sigma_{1i} = -\phi_{i,2}, \quad \sigma_{2i} = \phi_{i,1} \quad (2)$$

where the comma stands for differentiation with x_1 or x_2 . Hence

$$\phi_i(x_1, x_2) = \int^{x_1} \sigma_{2i}(\xi, x_2) d\xi$$

and the boundary conditions for $\sigma_{2i} = 0$ can be replaced by $\phi_i = 0$. In matrix notation u and ϕ , we may write (1a-1h) in one equation as

$$I_u u + I_\phi \phi = 0, \quad (3)$$

$$I_u + I_\phi = I. \quad (4)$$

In the above I is the 3x3 unit matrix and I_u, I_ϕ are 3x3 diagonal matrices whose diagonal elements are either one or zero. The special cases (1a) and (1h) correspond, respectively, to $I_u = 0, I_\phi = I$ and $I_u = I, I_\phi = 0$. Equation (3) encompasses all eight conditions in (1).

We will generalize the above derivation in two ways. First, the right hand side of (3) is replaced by something which is prescribed. Thus one may prescribe non-zero tractions or displacement at the boundary. Secondly, we replace (4) by

$$I_u^T I_u + I_\phi^T I_\phi = I, \quad (5)$$

$$I_u^T I_\phi = 0 = I_\phi^T I_u, \quad (6)$$

where I_u, I_ϕ are now subject to (5) and (6) but otherwise arbitrary, and the superscript T denotes the transpose. The I_u, I_ϕ defined earlier satisfy (5) and (6) but the new I_u, I_ϕ defined by (5), and (6) admit a wider class of boundary conditions. The explicit expressions of the new I_u, I_ϕ and the physical meanings of some of the I_u, I_ϕ will be presented. The modification required on the Stroh formalism to encompass the general boundary conditions is rather minimal. The final solution, which resembles the original unmodified solution, applies to a wide range of boundary conditions. One advantage of the generalized formalism is that, when a different boundary condition is desired, there is no need to re-formulate the problem.

REFERENCES

- [1] M. Z. Wang, T. C. T. Ting and Gongpu Yan, "The anisotropic elastic semi-infinite strip," Q. Appl. Math. In press.
- [2] A. N. Stroh, "Dislocations and cracks in anisotropic elasticity," Phil. Mag. 3, 625-646 (1958).
- [3] A. N. Stroh, "Steady state problems in anisotropic elasticity," J. Math. Phys. 41, 77-103 (1962).

Computation of Microstructure Utilizing Young Measure Representations*

R. A. Nicolaides[†] and Noel J. Walkington[‡]

Center for Nonlinear Analysis
Department of Mathematics, Carnegie Mellon University

Abstract

An algorithm is proposed for the solution of non-convex variational problems. In order to avoid representing highly oscillatory functions on a mesh, an associated Young measure, which characterizes such oscillations, is also approximated. Sample calculations demonstrate the viability of this approach.

keywords: Calculus of Variations, Young Measures.

1 Introduction

A recent development in continuum mechanics is the introduction of continuum energy functionals modeling nonlinear effects of crystal thermoelasticity [2, 6, 7, 8]. Among other things these functionals can be used to study displacive phase transformations and shape memory effects [9].

A characteristic feature of the energy functionals is their multiple well structure. Typically, each well represents a potential equilibrium state of the crystal, and at a transformation temperature more than one well is accessible to the crystal as a stable configuration.

The variational approach to finding an overall equilibrium state for the crystal requires that the energy functional be minimized in some suitable sense. In attempting such minimizations, one frequently encounters minimizing sequences of rapidly oscillating functions. These oscillations are usually a mathematical precursor to the formation of microstructure. This microstructure is characterized mathematically by probability distributions which, in principle, can be found by taking certain averages of the oscillatory functions.

In computational practice, the minimizing sequences are often constructed using a finite mesh, for example by finite elements. The oscillations referred to above then show up as grid scale oscillations of the (generally nonunique) minimizer. As the mesh is refined, the oscillations persist becoming more and more rapid while remaining of finite amplitude e.g. [4]. Usually, one

*This work was supported by the Army Research Office and National Science Foundation through the Center for Nonlinear Analysis

[†]Supported by Airforce Grant AFOSR F49620-92-J-0133.

[‡]Supported by National Science Foundation Grant No. DMS-9002768.

wishes to know the values of macroscopic quantities associated with the deformation. These are computed in two different ways. Essentially, linear functions of the deformation can be obtained as the limits of the same linear functions of the minimizing sequence. On the other hand, nonlinear functions of the deformation (including energy) in general have to be computed as expected values of the probability distribution mentioned in the previous paragraph.

In some situations, the probabilities that are needed for computing nonlinear functions of the deformation are known a priori. In this paper we are interested in the opposite case. Although in principle it must be possible to compute the probabilities from the oscillatory minimizing sequence, in practice this could be very difficult if there were a relatively large number of wells. Also, it is easy to imagine that a rather fine mesh would be necessary to accumulate enough data to permit the evaluation of stable averages. We refer to this as the "microscopic" approach.

An alternative method is to compute with the probabilities as dependent variables. However, this is feasible only if we have some information about the limiting probability distributions which can occur. It turns out that frequently there is enough prior information to permit the computation to be done in that way. Our main goal is to investigate this alternative approach, which we will call the "macroscopic" methodology. Its potential advantage is that since the probabilities are smoothly varying quantities, a relatively coarse mesh can be used to approximate them. In this way we can avoid the need to deal with the oscillations explicitly. Nonlinear functions of the deformation (and linear functions as a special case) may be approximated using the computed probability distributions.

In the sections which follow, we state our algorithm and explain the ideas behind it. Then we present the results of some model computations. The work reported is of a preliminary nature. So far, we do not have sufficient experience with the algorithm to make rational comparisons with other approaches. It is hoped to address these issues in a future report.

2 The macroscopic formulation

We will begin with a simplified presentation of some background information in variational methods which is sufficient for understanding the principles behind the formulation of the algorithm. More detailed accounts can be found in [5, 12, 13].

We consider variational integrals

$$J(u) := \int_{\Omega} F(x, u, \nabla u) dx, \quad u \in W_0^{1,p}(\Omega)^m, \quad (1)$$

where $\Omega \in R^n$ is a bounded domain and $F(\cdot, \cdot, \cdot)$ is continuous. Inhomogeneous boundary conditions can easily be accommodated if necessary. The case of most interest is when $F(x, u, \cdot)$ is not convex with respect to its last variable. The multiple well property of stored elastic energy functions causes this lack of convexity. In this case, the infimum of $J(\cdot)$ usually cannot be reached in $W_0^{1,p}(\Omega)^m$ and it is necessary to admit generalized solutions.

The standard example to illustrate this is

$$J(u) := \int_0^1 (u'^2 - 1)^2 + u^2 dx, \quad u \in W_0^{1,4}(0, 1).$$

The sequence $\{u_k\}$, whose first three members are illustrated in Figure 1b below, gives

$$J(u_k) = \int_0^1 u_k^2 dx \rightarrow 0 = \inf_{W_0^{1,4}} J(u),$$

so that $\{u_k\}$ is a minimizing sequence. Also clear is that $\lim_{k \rightarrow \infty} u_k = u \equiv 0$. On the other hand, $u'_k \not\rightarrow 0$ in any ordinary sense, and so for this $\{u_k\}$, $\inf_{W_0^{1,4}} J(u)$ is not attained. It is the existence of the two wells at ± 1 , illustrated in Figure 1a, which is responsible for the oscillatory behavior of the sequence $\{u'_k\}$.

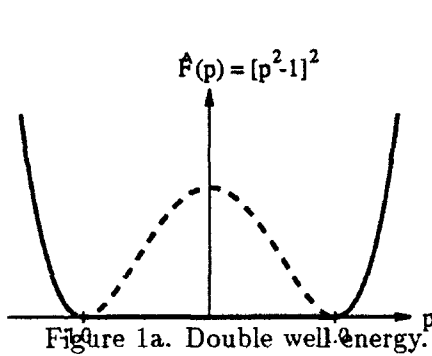


Figure 1a. Double well energy.

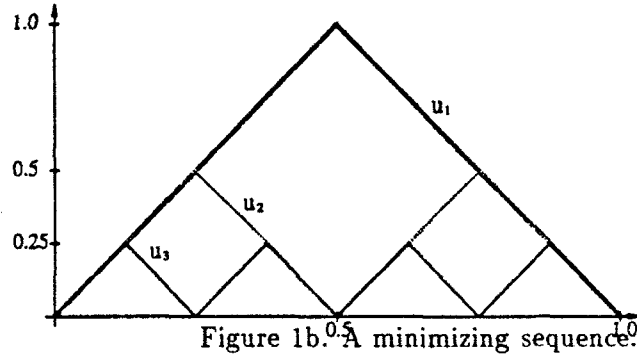


Figure 1b. A minimizing sequence.

The behavior of the sequence $\{u'_k\}$ strongly suggests that its "values" at any point $0 < x_0 < 1$ may be described by the probability distribution $u'(x_0) = \pm 1$ with probability $1/2$. In fact there is a general result from which this can be inferred: for any bounded sequence in $W^{1,p}(\Omega)^m$, $\|u_k\|_{1,p} \leq M$, $\{u_k\}$ contains a subsequence $\{u_{k_j}\}$ such that $u_{k_j} \rightharpoonup u \in L^p(\Omega)^m$. Additionally a subsequence of $\{u_{k_j}\}$ exists (denoted the same way) with the property that for any continuous g which is reasonably behaved at infinity, and for each $x \in \Omega$ there is a probability measure ν_x such that

$$g(\nabla u_{k_j}) \rightharpoonup G \in L^p(\Omega)^m, \quad (2)$$

where

$$G(x) = \int_{R^n} g(y) d\nu_x(y) \quad (3)$$

for almost all $x \in \Omega$. A useful version of this result, due to Kinderlehrer and Pedregal [10], states that if the sequence $\{u_k\}$ is a minimizing sequence for a variational problem having non-negative integrand with p -growth, then g may also have p -growth.

A family of probability measures $\{\nu_x\}$ obtained in this way is called a family of *gradient Young measures* [11]. Young measures also exist which are not gradient measures. They are derived in a similar way from bounded sequences in $L^p(\Omega)^m$.

There is a very useful characterization of Young measures ν_x due to Ball [1]. We will state this for gradient Young measures: let $\nu_{x,\delta}^k$ denote the probability distribution of the values of $\nabla u_k(z)$ as z is chosen uniformly at random from $B(x, \delta)$, the open ball with radius δ and center $x \in \Omega$. Then

$$\lim_{\delta \rightarrow 0} \lim_{k \rightarrow \infty} \left| \int_{R^n} g(y) d\nu_x(y) - \int_{R^n} g(y) d\nu_{x,\delta}^k(y) \right| = 0.$$

This result reveals how it is the minimizing sequence that determines the probability distribution ν_x and provides a way to approximate it.

The result (2)-(3) does not give any information about the structure of the measure ν_x , and in particular whether it is discrete. General results on this do not appear to be available. Nevertheless, there is a large class of problems where it is expected on physical grounds that ν_x is indeed discrete. This class includes most, if not all, of the continuum functionals used so far to model crystal energy. Since we want to make essential use of discreteness, we will introduce

it as a hypothesis. Specifically, we will assume that

$$\nu_x = \sum_{l=1}^L \lambda_l(x) \delta_{A_l(x)}, \quad (4)$$

$$\sum_{l=1}^L \lambda_l(x) = 1, \quad 0 \leq \lambda_l \leq 1, \quad (5)$$

where $\delta_{A_l(x)}$ denotes a Dirac mass with pole at $A_l(x)$ and $\lambda_l(x)$ varies measurably with x . References [3, 2, 8] contain examples satisfying the discreteness hypothesis.

Choosing g in (2) to be $F(x, u(x), \cdot)$ and denoting by $\{u_k\}$ a minimizing sequence bounded in $W^{1,p}(\Omega)^m$, we have

$$\lim_{j \rightarrow \infty} \int_{\Omega} F(x, u_k, \nabla u_k) dx = \int_{\Omega} \langle \nu_x, F(x, u(x), \cdot) \rangle dx,$$

where $\langle \nu_x, \cdot \rangle$ denotes the action on the right side of (3). Additionally, choosing g in (2)-(3) to be the identity mapping shows that

$$\nabla u := \sum_{l=1}^L \lambda_l(x) A_l(x).$$

These results motivate the following generalized variational problem: minimize

$$I(u) := \int_{\Omega} \langle \nu_x, F(x, u(x), \cdot) \rangle dx, \quad u \in W_0^{1,p}(\Omega)^m, \quad (6)$$

subject to

$$\nabla u(x) = \sum_{l=1}^L \lambda_l(x) A_l(x)$$

over suitable $A_l \in L^p(\Omega)^{mn}$, $\lambda_l \in L^\infty(\Omega)$, $l = 1, 2, \dots, L$. Solutions to this problem are regarded as generalized solutions to (1). Notice that classical solutions to (1) may be recovered from the generalized formulation by taking, say, $\lambda_1 = 1$.

The variables in the generalized formulation are, in principle, slowly varying or macroscopic.

3 Numerical Algorithm

In this section we consider discretizations of the generalized problem. Basically, we use continuous piecewise linear approximations for u , and piecewise constant approximations for the A_l and λ_l , $1 \leq l \leq L$. However, it is important to note that the A_l cannot be always be arbitrarily chosen, since the combination on the right of (4) must be a gradient Young measure. We present a general way to handle this issue.

3.1 Computing the Constraints

The algorithm presented above involves several constraints, namely,

$$\nabla u = \sum_{l=1}^L \lambda_l A_l, \quad \sum_{l=1}^L \lambda_l = 1, \quad \text{and} \quad 0 \leq \lambda_l \leq 1, \quad 1 \leq l \leq L$$

(recall that the discrete u is piecewise linear, so its gradient is piecewise constant, as are the discrete A_l and λ_l). In addition to these obvious constraints, when u is vector valued further constraints on the representation of the gradient are required to guarantee that $\nu = \sum_{l=1}^L \lambda_l \delta_{A_l}$ is a *gradient* Young measure. The constraints on $\{\lambda_l\}_{l=1}^L$ are convex and trivially accommodated; however, the constraints associated with the gradient are not convex. Moreover, since imposing constraints can be computationally taxing, it is imperative to resolve them in an efficient manner. Below we outline an algorithm that effectively eliminates the constraints on ∇u analytically.

We begin by considering the case with $L = 2$, i.e.

$$\nabla u = \lambda A_0 + (1 - \lambda) A_1.$$

Letting $b = A_1 - A_0$, we may write

$$A_0 = \nabla u - (1 - \lambda)b, \quad \text{and} \quad A_1 = \nabla u + \lambda b.$$

In this situation,

$$\begin{aligned} \langle F(x, u, \cdot), \nu \rangle &= \lambda F[x, u, A_0] + (1 - \lambda) F[x, u, A_1] \\ &= \lambda F[x, u, \nabla u - (1 - \lambda)b] + (1 - \lambda) F[x, u, \nabla u + \lambda b]. \end{aligned}$$

In the scalar case, $b \in R^n$ can be selected arbitrarily; however, when u is vector valued, $\nabla u \in R^{m \times n}$, and it is necessary and sufficient that $b = A_1 - A_0$ be a rank one matrix, $A_1 - A_0 = \mathbf{a} \otimes \mathbf{n}$, in order to obtain a gradient Young measure ($\mathbf{a} \in R^m$, $\mathbf{n} \in R^n$ may be chosen freely). i.e.

$$A_0 = \nabla u - (1 - \lambda)\mathbf{a} \otimes \mathbf{n}, \quad A_1 = \nabla u + \lambda\mathbf{a} \otimes \mathbf{n},$$

$$\begin{aligned} \langle F(x, u, \cdot), \nu \rangle &= \lambda F[x, u, A_0] + (1 - \lambda) F[x, u, A_1] \\ &= \lambda F[x, u, \nabla u - (1 - \lambda)\mathbf{a} \otimes \mathbf{n}] + (1 - \lambda) F[x, u, \nabla u + \lambda\mathbf{a} \otimes \mathbf{n}]. \end{aligned}$$

To obtain a representation of the gradient for arbitrarily large L , we repeat the construction as follows. Given A_0 and A_1 as above, write

$$A_0 = \lambda_0 A_{00} + (1 - \lambda_0) A_{01}, \quad A_1 = \lambda_1 A_{10} + (1 - \lambda_1) A_{11},$$

where

$$A_{01} - A_{00} = b_0, \quad \text{and} \quad A_{11} - A_{10} = b_1,$$

if u is scalar valued, and

$$A_{01} - A_{00} = \mathbf{a}_0 \otimes \mathbf{n}_0, \quad \text{and} \quad A_{11} - A_{10} = \mathbf{a}_1 \otimes \mathbf{n}_1,$$

if u is vector valued. The representation for the gradient then becomes,

$$\nabla u = \lambda \lambda_0 A_{00} + \lambda(1 - \lambda_0) A_{01} + (1 - \lambda) \lambda_1 A_{10} + (1 - \lambda)(1 - \lambda_1) A_{11}.$$

The quantities A_{00} etc. are determined from b (or \mathbf{a} and \mathbf{n}), λ , b_0 (or \mathbf{a}_0 and \mathbf{n}_0), λ_0 , etc., for example,

$$A_{01} = \nabla u - (1 - \lambda)b + \lambda_0 b_0$$

in the scalar case, and

$$A_{01} = \nabla u - (1 - \lambda)a \otimes n + \lambda_0 a_0 \otimes n_0$$

in the vector case.

By repeating this process N times, we obtain admissible Young measures consisting of 2^N Dirac masses. This construction is conveniently represented with a binary tree as shown in Figure 2. Each matrix occurring in the representation of the measure corresponds to a leaf on the tree, and is uniquely identified by a binary word of length N .

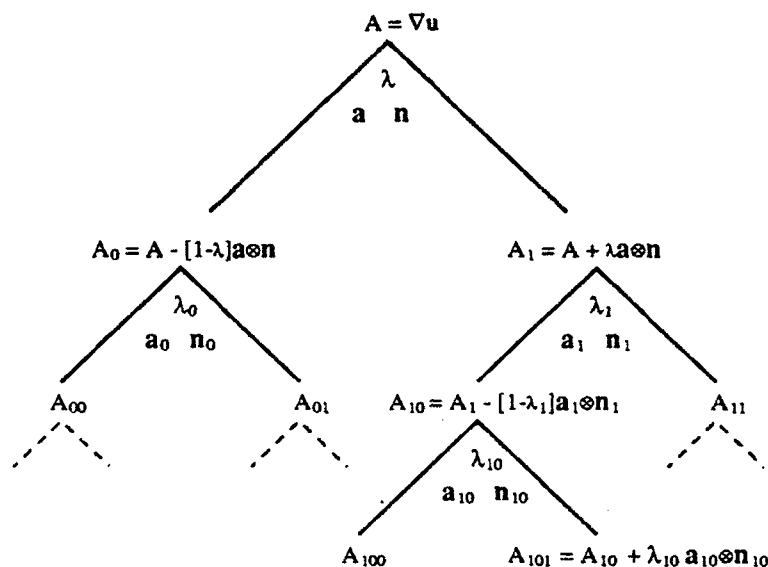


Figure 2. Binary tree representation of the micro structure

This representation of the gradient has the following desirable properties.

- If F is convex in its last variable, then trivially the minimum of $\langle F(x, u, \cdot), \nu \rangle$ is attained with $b = b_0 \dots \equiv 0$, $A_0 = A_1 = \dots \equiv \nabla u$, and in this situation the problem reduces to the classical algorithm for approximating the solution of elliptic problems using piecewise linear functions.
- Given a guess for the minimizing function u , minimizing with respect to the piecewise constant functions b , λ , b_0 etc. can be done in parallel over each element, suggesting the overhead associated with calculating a Young measure can be minimized by taking advantage of modern computer architectures.

4 Numerical Results

4.1 Computational Considerations

To obtain a solution of the discrete problem, simple relaxation was used in conjunction with the "numerical tricks" discussed below. The idea behind relaxation is to freeze all but one unknown, ξ (a nodal value of u , or a λ value for an element, etc.), and to make one Newton iteration for the Euler equation $dI/d\xi = 0$, i.e. $\xi^{n+1} = \xi^n - I'(\xi^n)/I''(\xi^n)$. The following embellishments were required for a practical algorithm.

- Clearly it is necessary to restrict λ to lie in $[0, 1]$. Moreover, since the algorithm degenerates when $\lambda = 0$, $\lambda = 1$, or $b = 0$, λ was required to satisfy $\epsilon \leq \lambda \leq 1 - \epsilon$ for some $\epsilon > 0$ (typically $\epsilon = 10^{-6}$ or 10^{-7}). Additionally, terms of the form $\epsilon(\lambda - 1/2)^2$ were added to the integrand to give a preferred value of $\lambda = 1/2$ when $b = 0$.
- Except for the Dirichlet data, initial values of $u = 0$, $b = 0$, and $\lambda = 1/2$ were chosen. It was observed that initially oscillations in u might develop before a suitable microstructure was found (this corresponds to computing a minimizing sequence directly). In order to suppress these oscillations while the microstructure developed, an "artificial viscosity" of the form $\mu(\Delta u)^2$ was added to the integrand. In all instances, μ was set to zero for the latter iterations.
- Since relaxation is a local algorithm, it is prone to "getting stuck" in local minima. It was frequently observed that microstructure would be present in one element, but not in an adjacent element. To remedy this problem, the micro-variables (λ , b etc.) were substituted for those in adjacent elements. If this lowered the energy, the modified microstructure was accepted. This non-local move was very effective for avoiding local minima.

This modified relaxation algorithm was found to be very effective for the computation of global minima. As with classical relaxation for the solution of elliptic problems, convergence was slow, especially in the latter iterates.

4.2 One Dimensional Examples

We consider one dimensional examples of the form.

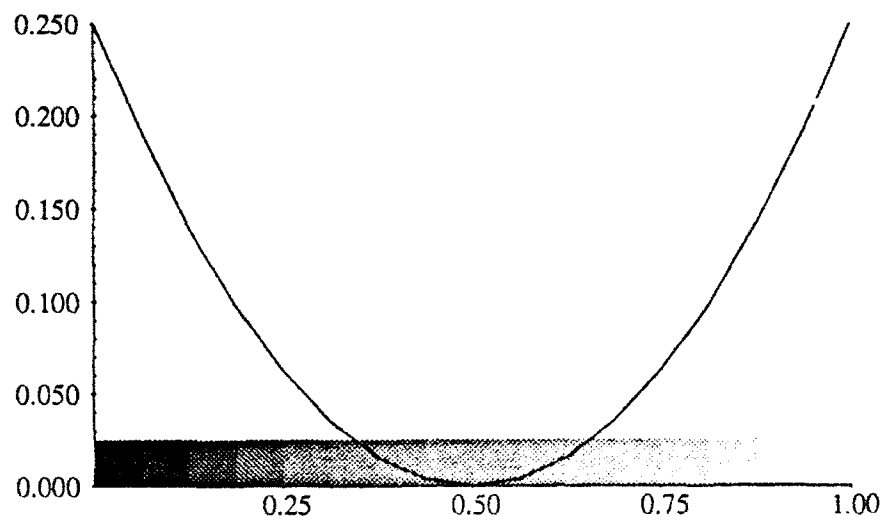
$$I(u) = \int_0^1 F(u') + (u - f)^2, \quad u(0) = u_0, \quad u(1) = u_1,$$

where $F(p) = (p^2 - 1)^2$ (see Figure 1) is the classical double well potential, and $f : [0, 1] \rightarrow \mathbb{R}$ is specified. For one dimensional problems, it suffices to consider only one level of the binary tree, i.e.

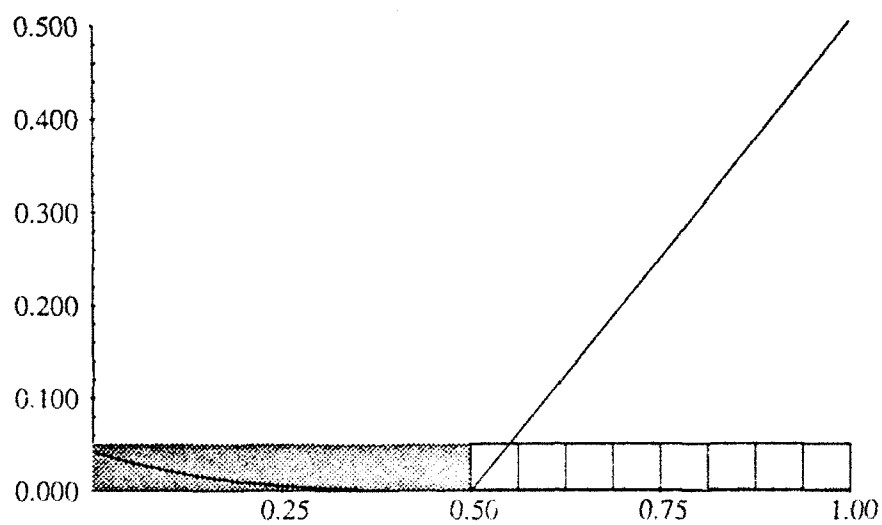
$$u' = \lambda A_0 + (1 - \lambda)A_1, \quad A_0 = u' - (1 - \lambda)b, \quad A_1 = u' + \lambda b.$$

4.2.1 Example 1 (Non-Homogeneous Young's Problem)

Setting $f(x) = (x - 1/2)^2$, the generalized solution of the variational problem is $u = f$, $\lambda = 1/2(1 + f')$, $b = 2$. For problems of this type (i.e. $-1 \leq f' \leq 1$), it is possible to show that the discrete solutions $\{u_h\}_{h>0}$ converge to u in $W^{1,p}(0,1)$ at the optimal rate of h . Similarly, $\lambda_h \rightarrow \lambda$ in $L^p(0,1)$ at optimal rate h . This is exhibited in Figure 3 where the $L^2(0,1)$ and $H^1(0,1)$ errors for u_h are tabulated. The solution obtained with a 16 element mesh is shown in Figure 4a.



(a) Example 1.



(b) Example 2.

Figure 4. One Dimensional Exapmles (16 Elements).

Black = well at -1, White = well at +1

No. Elements	Example 1		Example 2	
	$\ u - u_h\ _{L^2(0,1)}$	$\ u' - u'_h\ _{L^2(0,1)}$	$\ u - u_h\ _{L^2(0,1)}$	$\ u' - u'_h\ _{L^2(0,1)}$
4	0.010423	0.144338	0.002352	0.029793
8	0.002604	0.072121	0.000582	0.015258
16	0.000654	0.036115	0.000145	0.007673
32	0.000176	0.018454	0.000037	0.003867
$\ u\ _{L^2(0,1)}$ or $\ u'\ _{L^2(0,1)}$	0.111803	0.577350	0.205711	0.719047

Figure 3: Error Norms for One Dimensional Examples

4.2.2 Example 2

We consider a second less trivial example involving a "broken" extremal¹. The nonhomogeneous term is,

$$f(x) = -3/128(x - 1/2)^5 - 1/3(x - 1/2)^3,$$

and the solution, given by

$$u(x) = \begin{cases} f(x), & 0 \leq x \leq 1/2, \\ 1/24(x - 1/2)^3 + (x - 1/2), & 1/2 \leq x \leq 1, \end{cases}$$

has microstructure in $(0, 1/2)$ and is "elliptic" on $(1/2, 1)$. On $(0, 1/2)$, $\lambda = 1/2(1 + u')$ and $b = 2$. Note that the derivative of u jumps from zero to one at $x = 1/2$. Figure 3 exhibits the optimal rates of convergence observed for $\{u_h\}_{h>0}$ in $L^2(0, 1)$ and $H^1(0, 1)$. The solution for a 16 element mesh is shown in Figure 4b.

4.3 Two Dimensional Example

We consider examples of the form

$$I(u) = \int_{\Omega} |\nabla u - \mathbf{w}_1|^2 |\nabla u - \mathbf{w}_2|^2,$$

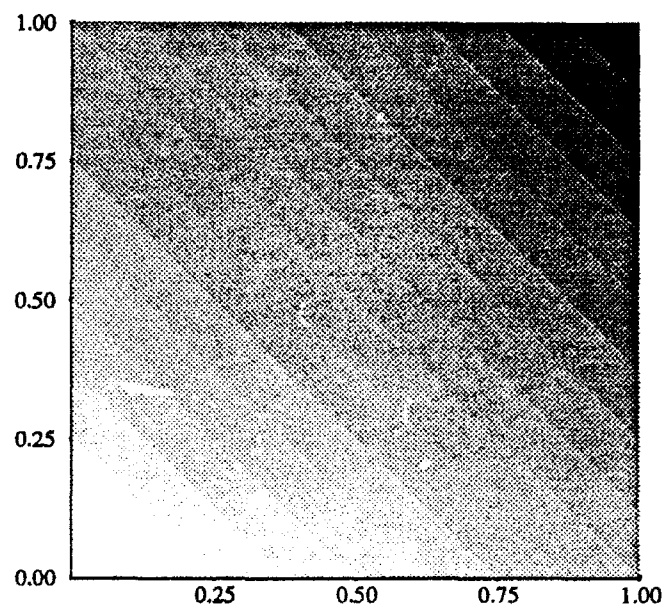
where $\mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^2$ are the locations of the energy wells. We chose $\Omega = (0, 1)^2$ to be the unit square, and impose Dirichlet boundary conditions on u . Triangular meshes are constructed by dividing the region into similar squares, and dividing them in two along the diagonal with slope -1 . We consider examples where the slope of u lies on the line joining \mathbf{w}_1 and \mathbf{w}_2 , so that the micro-structure can be represented by a gray scale with \mathbf{w}_1 colored black and \mathbf{w}_2 colored white.

We present two examples, one being obtained from the other by a rotation of 90° . This illustrates what happens when the mesh is most favorably and least favorably aligned with the contours of the exact solution, u . In Figure 5a, a solution with $\mathbf{w}_1 = (-1, -1)$ and $\mathbf{w}_2 = (1, 1)$ is shown having exact solution

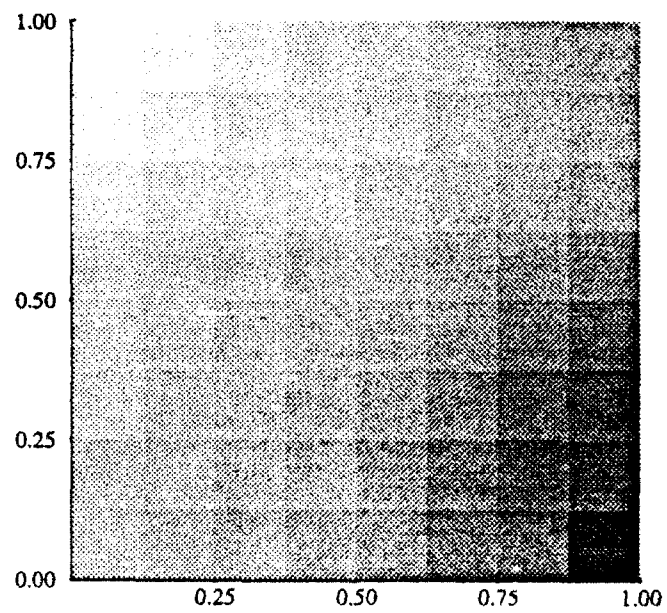
$$u(x, y) = \frac{2}{e^2 - 1} e^{(x+y)} - \frac{e^2 + 1}{e^2 - 1} (x + y),$$

$$\nabla u = \lambda \mathbf{w}_1 + (1 - \lambda) \mathbf{w}_2, \quad \lambda(x, y) = (e^2 - e^{x+y}) / (e^2 - 1).$$

¹This solution was suggested by Luc Tartar.



(a) Black = (1,1) well, White = (-1,-1) well.



(b) Black = (1,-1) well, White = (-1,1) well.

Figure 5. Two Dimensional Example, 8 x 8 Mesh.

The solution corresponding to a 90° rotation is shown in Figure 5b. Here $\mathbf{w}_1 = (1, -1)$, $\mathbf{w}_2 = (-1, 1)$ with exact solution

$$u(x, y) = \frac{2e}{e^2 - 1} e^{(x-y)} - \frac{e^2 + 1}{e^2 - 1} (x - y),$$

$$\nabla u = \lambda \mathbf{w}_1 + (1 - \lambda) \mathbf{w}_2, \quad \lambda(x, y) = (e^{1+x-y} - 1)/(e^2 - 1).$$

5 Concluding Remarks

In conclusion, the computations show that the overall approach is a useful one, and that it does produce optimal rates of convergence under mesh refinement. Certainly, more work must be done to implement the vector case, and also to improve the performance of algebraic solvers. It is hoped to address these matters in the future.

Acknowledgment: We thank David Kinderlehrer for many valuable conversations and suggestions.

References

- [1] J. M. Ball. A version of the fundamental theorem for young measures. In D. Serre, editor, *Partial Differential Equations and Continuum Models of Phase Transitions*. Springer, May 1988.
- [2] J. M. Ball and R. D. James. Fine phase mixtures as minimizers of energy. *Archive for Rational Mechanics and Analysis*, 100:13-52, 1987.
- [3] K. Bhattacharya. Wedge-like microstructures in martensites. *Acta Metall. Mater.*, 39(10):2431-2444, 1991.
- [4] C. Collins and M. Luskin. The computation of the austenitic-martensitic phase transition. In M. Rasche, D. Serre, and M. Slemrod, editors, *Partial Differential Equations and Continuum Models of Phase Transitions, Lecture Notes in Physics 344*, pages 34-50. Springer Verlag, 1989.
- [5] B. Dacorogna. *Direct Methods in the Calculus of Variations*. Springer Verlag, 1989.
- [6] J. L. Ericksen. Stable equilibrium configurations of elastic crystals. *Archive for Rational Mechanics and Analysis*, 94:1-14, 1986.
- [7] J. L. Ericksen. Some constrained elastic crystals. In J. M. Ball, editor, *Material Instabilities in Continuum Mechanics and Related Mathematical Problems*, pages 119-135. Oxford University Press, May 1988.
- [8] R. D. James and D. Kinderlehrer. Theory of diffusionless phase transitions. In M. Rasche, D. Serre, and M. Slemrod, editors, *Partial Differential Equations and Continuum Models of Phase Transitions, Lecture Notes in Physics 344*, pages 51-84. Springer Verlag, 1989.
- [9] R. D. James and D. Kinderlehrer. Frustration in ferromagnetic materials. *Continuum Mechanics and Thermodynamics*, 2:215-239, 1990.

- [10] D. Kinderlehrer and P. Pedregal. Weak convergence of integrands and the young measure representation. Technical Report 90-87-NAMS-3, Carnegie Mellon University, Aug. 1990.
- [11] D. Kinderlehrer and P. Pedregal. Characterization of young measures generated by gradients. *Archive for Rational Mechanics and Analysis*, Preprint.
- [12] C. B. Morrey. *Multiple Integrals in the Calculus of Variations*. Springer, 1966.
- [13] L. C. Young. *Lectures on the Calculus of Variations and Optimal Control*. Chelsea, 1980.

Kinetically Driven Elastic Phase Boundary Motion Activated by Concurrent Dynamic Pulses¹

Jiehliang Lin and Thomas J. Pence

Department of Materials Science and Mechanics

Michigan State University,

East Lansing, MI 48824-1226

Abstract: We consider the behavior of a phase boundary that is subjected to concurrent dynamic pulses, one from each side, in the event that the phase boundary motion is governed by a simple kinetic relation. The total energy loss is contrasted to that which would occur if the two pulses were not concurrent.

1. Introduction

In one spatial dimension, nonlinearly elastic stress-strain laws that are not monotonic provide a model for stress induced diffusionless phase transformations [1975E]. In this setting the purely elastic interaction of acoustic pulses with phase boundaries are mathematically underdetermined, necessitating the consideration of additional physical effects. Possibilities include criteria which capture kinetic effects [1987T], [1990G], [1991A], [1991F], impedance effects [1991P], dissipative effects [1980J], [1986H], [1991PP], [1992P] or other phenomena not accounted for by the purely elastic theory [1983H], [1991T].

In [1992L] we considered maximally dissipative dynamical motion, meaning that phase boundaries move so as to maximize the instantaneous local rate at which purely mechanical energy is converted to nonmechanical energy. Understanding acoustic pulse reverberation processes governed by this principle focuses attention on a concurrent pulse problem, namely a situation in which two pulses, one from each side, act simultaneously on a previously stationary phase boundary. In particular, [1992L] addressed the question:

How does the total energy loss for a concurrent pulse problem governed by the maximum dissipation rate criterion (M.D.C.) compare to the combined energy loss for two subsidiary problems: one involving only the pulse which impinges from the front (governed by M.D.C.), and the other involving only the pulse which impinges from the back (also governed by M.D.C.)?

We showed in [1992L] that the answer to this question was dependent on whether the two pulses were of the same sign with respect to the strains in the ambient initial static state, or whether instead they were of opposite sign. Namely, we concluded, under the maximum dissipation rate criterion (M.D.C.), that the concurrent pulse encounter suffers the greater energy loss in the event that both incoming pulses are of the same (strain) sign, whereas the concurrent pulse encounter suffers the lesser energy loss in the event that the incoming pulses are of opposite sign.

Our purpose here is to consider the same question for the case in which the interaction

1. Supported by the U.S. Army Research Office under contract DAAL03-89-G-0089.

dynamics are not governed by the maximum dissipation rate criterion, but are instead governed by a *linear kinetic relation* (L.K.R.) between the driving traction on the phase boundary and the phase boundary velocity. This linear kinetic relation is given in (3.4). Thus we address the question

How does the total energy loss for a concurrent pulse problem governed by a linear kinetic relation (L.K.R.) compare to the combined energy loss for two subsidiary problems: one involving only the pulse which impinges from the front (governed by L.K.R.), and the other involving only the pulse which impinges from the back (also governed by L.K.R.)?

We show that the latter is greater than the former if both incoming pulses are of opposite sign, which is like the result obtained in [1992L] for the M.D.C. case. In the event that the incoming pulses are of the same sign, and if each pulse is sufficiently small, we find that the former is greater than the latter. However we also find numerically that certain situations involving sufficiently large pulses of the same sign result in the concurrent pulse encounter suffering the lesser energy loss.

2. Families of Solutions to the Concurrent Pulse Problem

Since the framework for the problem addressed here does not depart from that addressed in [1992L] until after the invocation of the particular resolution of the nonuniqueness issue, it follows that the families of solutions to the concurrent pulse problem here are identical to those obtained in [1992L]. Accordingly, we here repeat, verbatim for completeness, the derivation of these solution families as previously given in [1992L]:

Let τ , γ and v denote respectively stress, strain and particle velocity. Following [1991P], we consider a layer, $0 < x < h$, composed of an elastic material whose stress-strain behavior in one dimension is given by

$$\tau = \hat{\tau}(\gamma) \equiv \begin{cases} c^2 \gamma & \text{for } 0 \leq \gamma \leq \gamma_M \\ \hat{\tau}_u(\gamma) & \text{for } \gamma_M \leq \gamma \leq \gamma_m \\ c^2 \gamma + d & \text{for } \gamma \geq \gamma_m \end{cases}, \quad \hat{\tau}(-\gamma) = -\hat{\tau}(\gamma), \quad (2.1)$$

where c and d are constants and $\hat{\tau}_u(\gamma)$ is a smooth decreasing function that renders $\hat{\tau}(\gamma)$ continuous. The layer is assumed to be initially pre-stressed in equilibrium, so that $v=0$, with a single phase boundary at $x = s_0$ separating high strain phase with $\gamma=\gamma_b$ in $x < s_0$ from low strain phase with $\gamma=\gamma_a$ in $x > s_0$. The strain values γ_a and γ_b are taken to be the well known Maxwell strains which have the geometrical interpretation of cutting off equal areas on the stress strain curve (Fig. 1). An immediate consequence is that the initial configuration is one of minimum energy [1975E] for the prevailing boundary conditions governing the initial equilibrium configuration. Any subsequent change in the boundary conditions will give rise to changes in the strain and velocity fields governed by the equations²

$$v_x - \gamma_t = 0, \quad \hat{\tau}'(\gamma) \gamma_x - v_t = 0. \quad (2.2)$$

In particular, we consider a loading condition at $x=0$ that gives rise to a square wave pulse with

2. Primes and subscripts denote differentiation in the usual fashion. Note also that we have taken the density to be equal to one in (2.2)₂.

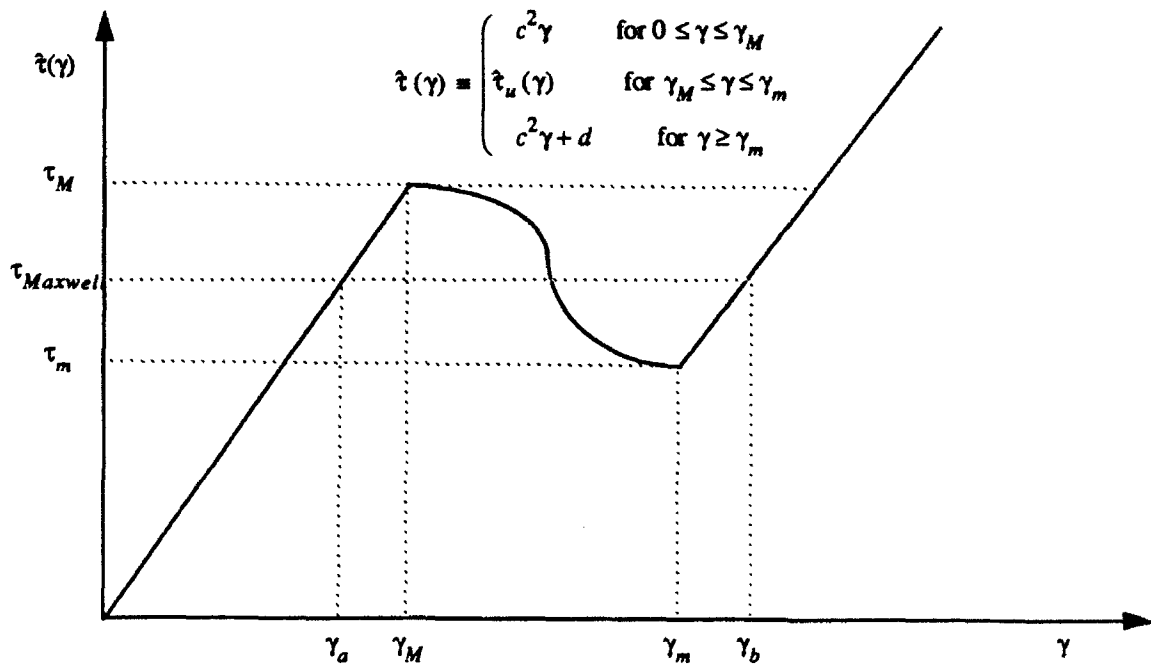


Fig. 1. Stress-strain constitutive response as described by (2.1). The descending portion of such a constitutive response function is associated with unstable material behavior [1975E]. The strain intervals: $[-\gamma_M, \gamma_M]$, $[\gamma_M, \gamma_m]$, $[\gamma_m, \infty)$, correspond respectively to a low strain phase, an unstable phase and a high strain phase.

strain $\gamma_b + \Delta\gamma_1$ over a time interval t_b , and a loading condition at $x = h$ that gives rise to a square wave pulse with strain $\gamma_a + \Delta\gamma_2$ over a time interval t_a . We shall not concern ourselves with the specific loading conditions needed to generate these pulses, nor with restrictions upon $\Delta\gamma_1$ and $\Delta\gamma_2$ necessary to ensure compatibility with (2.1) other than to note that these issues can be treated in a systematic fashion [1991P]. According to (2.2), each pulse will travel toward the phase boundary with speed c ; furthermore the right moving pulse has width ct_b and particle velocity given by $-c\Delta\gamma_1$, while the left moving pulse has width ct_a and particle velocity given by $c\Delta\gamma_2$. The encounter of such a right moving pulse with the phase boundary is treated in [1991P] on the assumption that the encounter ends before the arrival of any pulse from the other side. Our purpose here, however, is to study such a concurrent encounter. There are four generic cases: (rr) , (rl) , (lr) , (ll) , where (rr) denotes the case where the *right moving* pulse (with strain increment $\Delta\gamma_1$) encounters the phase boundary first and also terminates last, (rl) denotes the case where the *right moving* pulse encounters the phase boundary first, but the encounter with the *left moving* pulse terminates last, and the remaining two cases are defined accordingly. For the remainder of this section, and also for Section 3, we shall restrict attention to the (rl) case. There are then three distinct interaction periods: Π_1 in which only the right moving pulse encounters the phase boundary, Π_{con} in which both pulses encounter the phase boundary concurrently, and Π_2 in which only the left moving pulse encounters the phase boundary. Figure 2 diagrams these encounters in the (x, t) -plane. According to this figure, the following additional assumptions are also implicit in our treatment: (A1) the phase boundary remains at rest unless acted on by a pulse, (A2) phase transitions take place only by movement of the pre-existing phase boundary, and (A3) the phase boundary

velocity is constant during each of the three interaction periods and these three phase boundary velocities obey

$$-c < \dot{s}_1 < c, \quad -c < \dot{s}_{con} < c, \quad -c < \dot{s}_2 < c. \quad (2.3)$$

Further discussion of these issues can be found in [1991P]. In addition we have depicted the phase boundary as coming to rest after the complete encounter has ended, in which case the fields return to their initial conditions on each side of the since displaced phase boundary.

In Figure 2, the subscripts $T1$ and $R1$ denote the fields in the transmitted and reflected pulses associated with interaction period Π_1 . In addition, the (x,t) -domain with combined incoming and reflected pulse during the interaction period Π_1 is denoted by subscript $S1$. A similar convention is followed for subscripts $T2$, $R2$, and $S2$ for the pulses associated with interaction period Π_2 . The fields associated with the combination of $T1$ and the incoming pulse characterized by $\Delta\gamma_2$ is denoted by subscript $T1i2$. Finally, there are four additional (x,t) -domains associated with pulses that arise as a consequence of the concurrent interaction period Π_{con} ; these are denoted by the subscripts $S1T2$, $R1T2$, $T1S2$, and $T1R2$. A consequence of (A3), (2.1) and (2.2) is that the value of strain and particle velocity are individually constant on the individual (x,t) -domains associated with the 11 symbols $T1$, $R1$, $S1$, $T1i2$, $T2$, $R2$, $S2$, $S1T2$, $R1T2$, $T1S2$, and $T1R2$. The correspond-

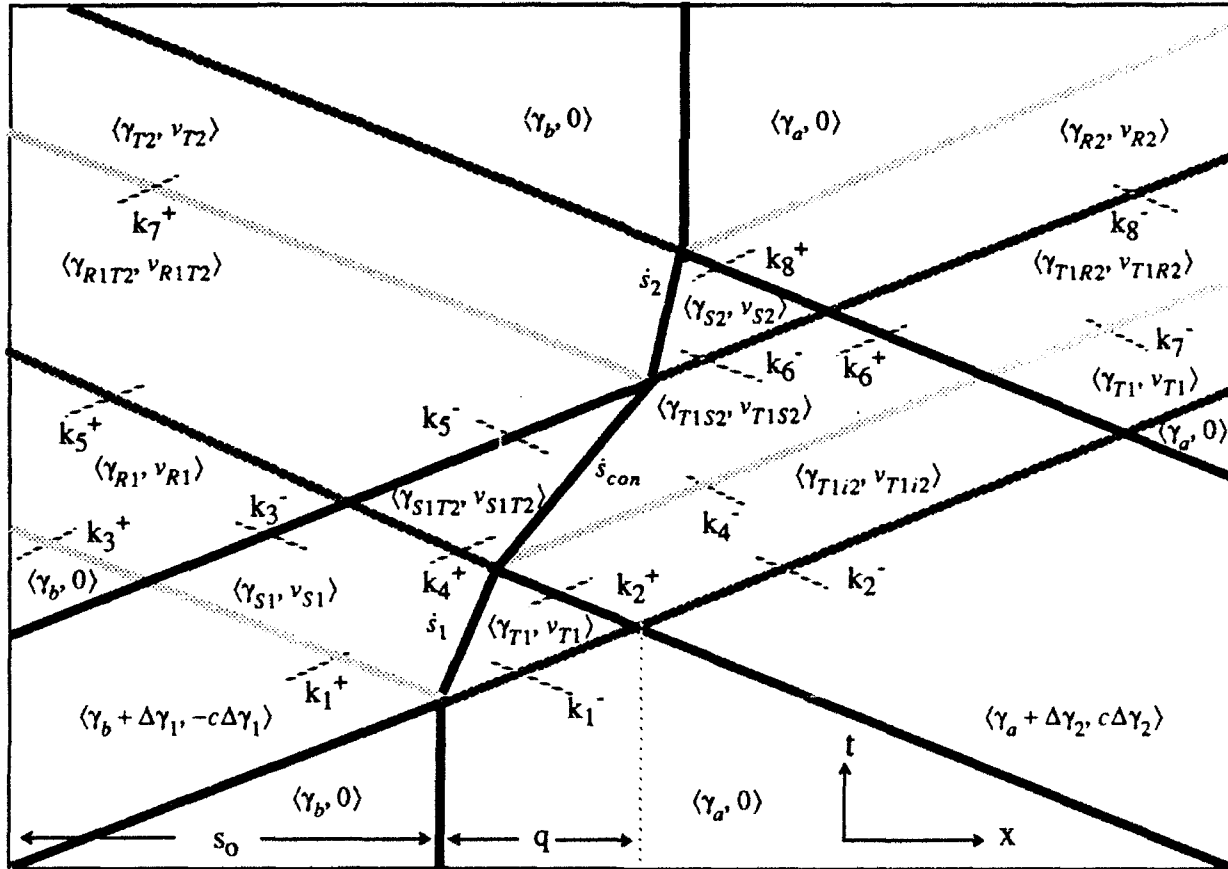


Fig. 2. Concurrent encounter of a right moving shear pulse, a left moving shear pulse and a phase boundary. The shear strain γ and velocity v in these incoming pulses and the generated pulses are denoted by $\langle \gamma, v \rangle$. The characteristic curves are indicated by dashed line segments.

ing 22 unknown values for strain and velocity, in conjunction with the three unknown phase boundary velocities \dot{s}_1 , \dot{s}_{con} , and \dot{s}_2 , comprise the unknown quantities in the complete encounter problem. Relations connecting these 25 unknown values to the parameters c , γ_b , γ_a , $\Delta\gamma_1$, $\Delta\gamma_2$ which characterize the material, initial conditions, and loading conditions follow from the theory of Riemann invariants as applied to (2.1), (2.2). In particular, this gives that $v-c\gamma$ is constant on any line segment with slope $\frac{dx}{dt}=c$ in the (x,t) -plane provided that it does not cross the phase boundary. Similarly $v+c\gamma$ is constant on all line segments with slope $\frac{dx}{dt}=-c$ that do not cross the phase boundary. Each of these Riemann invariant conditions generates 8 algebraic equations relating $\{\gamma, v\}$ pairs between contiguous (x,t) -domains; the associated connecting line segments are denoted by K_I^+, \dots, K_8^+ and K_I^-, \dots, K_8^- in Figure 2. Across the phase boundary, the two Rankine-Hugoniot conditions

$$[[v]] = -\dot{s}[[\gamma]], \quad [[\tau]] = -\dot{s}[[v]], \quad (2.4)$$

associated with (2.1), (2.2) are required to hold. These give rise to an additional 6 algebraic equations, 2 for each of the 3 interaction periods Π_1 , Π_{con} , and Π_2 . Thus in total there are 22 equations relating the 25 unknown values. Regarding the three phase boundary velocities as parameters, the 22 equations are linear in the 22 strain and particle velocities. The resulting 22x22 coefficient matrix is nonsingular provided that none of the three phase boundary velocities \dot{s}_1 , \dot{s}_{con} , and \dot{s}_2 , take on the values c or $-c$. Hence (2.3) ensures that the system can be inverted. Certain simplifications are achieved in the resulting problem due to various uncouplings (i.e. zero blocks in the coefficient matrix). For example $\{\gamma_{S1}, v_{S1}\}$ and $\{\gamma_{T1}, v_{T1}\}$ can be found from the 2 Riemann invariant conditions associated with K_I^+ and K_I^- , along with the 2 Rankine-Hugoniot conditions associated with interaction period Π_1 . The resulting 22 field quantities are thus found to be given by:

$$\begin{aligned} \gamma_{S1} &= \gamma_b + \Delta\gamma_1 - \frac{(\gamma_b - \gamma_a)\dot{s}_1}{2(\dot{s}_1 + c)}, & v_{S1} &= -c\Delta\gamma_1 - \frac{c(\gamma_b - \gamma_a)\dot{s}_1}{2(\dot{s}_1 + c)}, \\ \gamma_{R1} &= \gamma_b - \frac{(\gamma_b - \gamma_a)\dot{s}_1}{2(\dot{s}_1 + c)}, & v_{R1} &= -\frac{c(\gamma_b - \gamma_a)\dot{s}_1}{2(\dot{s}_1 + c)}, \\ \gamma_{T1} &= \gamma_a + \Delta\gamma_1 + \frac{(\gamma_b - \gamma_a)\dot{s}_1}{2(\dot{s}_1 - c)}, & v_{T1} &= -c\Delta\gamma_1 - \frac{c(\gamma_b - \gamma_a)\dot{s}_1}{2(\dot{s}_1 - c)}, \\ \gamma_{T1i2} &= \gamma_a + \Delta\gamma_1 + \Delta\gamma_2 + \frac{(\gamma_b - \gamma_a)\dot{s}_1}{2(\dot{s}_1 - c)}, & v_{T1i2} &= -c\Delta\gamma_1 + c\Delta\gamma_2 - \frac{c(\gamma_b - \gamma_a)\dot{s}_1}{2(\dot{s}_1 - c)}, \\ \gamma_{S1T2} &= \gamma_b + \Delta\gamma_1 + \Delta\gamma_2 - \frac{(\gamma_b - \gamma_a)\dot{s}_{con}}{2(\dot{s}_{con} + c)}, & v_{S1T2} &= -c\Delta\gamma_1 + c\Delta\gamma_2 - \frac{c(\gamma_b - \gamma_a)\dot{s}_{con}}{2(\dot{s}_{con} + c)}, \\ \gamma_{T1S2} &= \gamma_a + \Delta\gamma_1 + \Delta\gamma_2 + \frac{(\gamma_b - \gamma_a)\dot{s}_{con}}{2(\dot{s}_{con} - c)}, & v_{T1S2} &= -c\Delta\gamma_1 + c\Delta\gamma_2 - \frac{c(\gamma_b - \gamma_a)\dot{s}_{con}}{2(\dot{s}_{con} - c)}, \\ \gamma_{R1T2} &= \gamma_b + \Delta\gamma_2 - \frac{(\gamma_b - \gamma_a)\dot{s}_{con}}{2(\dot{s}_{con} + c)}, & v_{R1T2} &= c\Delta\gamma_2 - \frac{c(\gamma_b - \gamma_a)\dot{s}_{con}}{2(\dot{s}_{con} + c)}, \end{aligned}$$

$$\begin{aligned}
\gamma_{T1R2} &= \gamma_a + \Delta \gamma_1 + \frac{(\gamma_b - \gamma_a) \dot{s}_{con}}{2(\dot{s}_{con} - c)}, & v_{T1R2} &= -c \Delta \gamma_1 - \frac{c(\gamma_b - \gamma_a) \dot{s}_{con}}{2(\dot{s}_{con} - c)}, \\
\gamma_{S2} &= \gamma_a + \Delta \gamma_2 + \frac{(\gamma_b - \gamma_a) \dot{s}_2}{2(\dot{s}_2 - c)}, & v_{S2} &= c \Delta \gamma_2 - \frac{c(\gamma_b - \gamma_a) \dot{s}_2}{2(\dot{s}_2 - c)}, \\
\gamma_{R2} &= \gamma_a + \frac{(\gamma_b - \gamma_a) \dot{s}_2}{2(\dot{s}_2 - c)}, & v_{R2} &= -\frac{c(\gamma_b - \gamma_a) \dot{s}_2}{2(\dot{s}_2 - c)}, \\
\gamma_{T2} &= \gamma_b + \Delta \gamma_2 - \frac{(\gamma_b - \gamma_a) \dot{s}_2}{2(\dot{s}_2 + c)}, & v_{T2} &= c \Delta \gamma_2 - \frac{c(\gamma_b - \gamma_a) \dot{s}_2}{2(\dot{s}_2 + c)}.
\end{aligned} \tag{2.5}$$

The phase boundary velocities \dot{s}_1 , \dot{s}_{con} , and \dot{s}_2 , are undetermined by the above procedure and can be regarded as parametrizing all possible solutions. In addition to (2.3), various additional restrictions upon the phase boundary velocities will arise due to the requirement that the strain values in (2.5) remain confined to the intervals associated with the different branches of the stress-strain relation (2.1). The net effect of these considerations is to generate additional inequality constraints beyond (2.3) on the phase boundary velocities. The totality of inequality constraints are not mutually exclusive provided that $\Delta \gamma_1$ and $\Delta \gamma_2$ are sufficiently small. If, however, $\Delta \gamma_1$ and $\Delta \gamma_2$ are large, then mutual exclusivity may prevail (see [1991P]). We shall henceforth assume that we are dealing with values of $\Delta \gamma_1$ and $\Delta \gamma_2$ which do not give rise to mutual exclusivity so that three non-empty parametrization intervals, \mathfrak{I}_1 , \mathfrak{I}_{con} , and \mathfrak{I}_2 , exist for the three phase boundary velocities.

3. Solutions Obeying a Linear Kinetic Relation for the (rl)-case

As mentioned in the Introduction, the freedom to determine the phase boundary velocity allows the theory to accommodate additional requirements upon conditions which govern phase boundary motion. It is at this juncture that the treatment given here departs from that given in [1992L] which considered the maximum dissipation criterion (M.D.C.) as the method for resolving the nonuniqueness issue. Here we instead examine a different operative condition, namely a linear kinetic relation (L.K.R.). Recall that the motion of a phase boundary gives rise to a change in the total mechanical energy stored in the mechanical fields [1980J]. In particular, if $\gamma^+ = \gamma(s(t)^+, t)$ and $\gamma^- = \gamma(s(t)^-, t)$ are the strains directly adjacent to the phase boundary, then the energy loss rate, or dissipation rate, is given by

$$D(t) = \dot{s}(t)f(t), \quad f(t) = \left(\int_{\gamma^-}^{\gamma^+} \tilde{\tau}(\gamma) d\gamma - \frac{1}{2} (\tilde{\tau}(\gamma^+) + \tilde{\tau}(\gamma^-)) (\gamma^+ - \gamma^-) \right) \tag{3.1}$$

where $f(t)$ can be interpreted as the phase boundary traction [1991A]. For the concurrent pulse problem, with strains as given in (2.5), one finds that the dissipation rate and phase boundary driving tractions during the three interaction periods are given by

$$\begin{aligned}
D_1 &= -\frac{1}{2\dot{s}_1} \{ c^2 (c^2 - \dot{s}_1^2) \{ (\gamma_{T1} - \gamma_a)^2 - (\gamma_{S1} - \gamma_b)^2 \} \}, & f_1 &= \frac{D_1}{\dot{s}_1}, \\
D_{con} &= -\frac{1}{2\dot{s}_{con}} \{ c^2 (c^2 - \dot{s}_{con}^2) \{ (\gamma_{T1S2} - \gamma_a)^2 - (\gamma_{S1T2} - \gamma_b)^2 \} \}, & f_{con} &= \frac{D_{con}}{\dot{s}_{con}}, \\
D_2 &= -\frac{1}{2\dot{s}_2} \{ c^2 (c^2 - \dot{s}_2^2) \{ (\gamma_{S2} - \gamma_a)^2 - (\gamma_{T2} - \gamma_b)^2 \} \}, & f_2 &= \frac{D_2}{\dot{s}_2}.
\end{aligned} \tag{3.2}$$

where use has been made of the special equal area property of the Maxwell strains γ_a and γ_b which characterize the initial configuration.

A kinetic relation is now assumed to govern the motion of the phase boundary. As discussed in [1991A], one form for such a kinetic relation is an additional relation between phase boundary velocity \dot{s} and the phase boundary driving traction f :

$$\dot{s} = F(f) \quad (3.3)$$

where the new constitutive function $F(f)$ is required to obey $f \cdot F(f) \geq 0$ in order to deliver $D(t) \geq 0$ and hence ensure compatibility with a purely mechanical version of the second law of thermodynamics that is appropriate in the present setting. In this study we shall consider the case of a linear kinetic relation $F(f) = kf$ where the phase boundary mobility k is a positive constant so that (3.3) becomes

$$\dot{s} = kf \quad (k > 0). \quad (3.4)$$

Hence, entering (3.4) with (3.2) and (2.5) one obtains the following implicit equations for \dot{s}_1 , \dot{s}_{con} , and \dot{s}_2 :

$$\begin{aligned} \Delta\gamma_1 &= \frac{\dot{s}_1}{kc^2(\gamma_b - \gamma_a)} + \frac{(\gamma_b - \gamma_a)c\dot{s}_1}{2(c^2 - \dot{s}_1^2)}, \\ \Delta\gamma_1 + \Delta\gamma_2 &= \frac{\dot{s}_{con}}{kc^2(\gamma_b - \gamma_a)} + \frac{(\gamma_b - \gamma_a)c\dot{s}_{con}}{2(c^2 - \dot{s}_{con}^2)}, \\ \Delta\gamma_2 &= \frac{\dot{s}_2}{kc^2(\gamma_b - \gamma_a)} + \frac{(\gamma_b - \gamma_a)c\dot{s}_2}{2(c^2 - \dot{s}_2^2)}. \end{aligned} \quad (3.5)$$

Each of the equations (3.5) admits a unique solution obeying (2.3) which we shall denote by $\dot{s}_1^{(kr)}$, $\dot{s}_{con}^{(kr)}$, and $\dot{s}_2^{(kr)}$. For a given pair $(\Delta\gamma_1, \Delta\gamma_2)$ it may or may not be the case that $\dot{s}_1^{(kr)} \in \mathfrak{I}_1$, $\dot{s}_{con}^{(kr)} \in \mathfrak{I}_{con}$, and $\dot{s}_2^{(kr)} \in \mathfrak{I}_2$, however in what follows we shall assume that these inclusions hold. Thus we deal with a set in the $(\Delta\gamma_1, \Delta\gamma_2)$ -plane which ensure that $\dot{s}_1^{(kr)} \in \mathfrak{I}_1$, $\dot{s}_{con}^{(kr)} \in \mathfrak{I}_{con}$, and $\dot{s}_2^{(kr)} \in \mathfrak{I}_2$ for the remainder of this communication, even though we do not here investigate the extent of this set.

We now introduce normalized phase boundary velocities, normalized pulse strain increments and normalized mobility k as follows:

$$\bar{\dot{s}} = \frac{\dot{s}}{c}, \quad \Delta\bar{\gamma} = \frac{\Delta\gamma}{(\gamma_b - \gamma_a)}, \quad \bar{k} = k(\gamma_b - \gamma_a)^2 c \quad (3.6)$$

where subscripted and superscripted quantities such as $\dot{s}_{con}^{(kr)}$ are defined in the obvious fashion by these same normalizations. Note that $-1 < \bar{\dot{s}} < 1$. Then the associated values for D_1 , D_{con} , and D_2 , which will be denoted by $D_1^{(kr)}$, $D_{con}^{(kr)}$, and $D_2^{(kr)}$, are given by

$$\begin{aligned}
D_1^{(kr)} &= \frac{(\dot{s}_1^{(kr)})^2}{k} = \frac{c^3 (\gamma_b - \gamma_a)^2}{\bar{k}} (\dot{s}_1^{(kr)})^2, \\
D_{con}^{(kr)} &= \frac{(\dot{s}_{con}^{(kr)})^2}{k} = \frac{c^3 (\gamma_b - \gamma_a)^2}{\bar{k}} (\dot{s}_{con}^{(kr)})^2, \\
D_2^{(kr)} &= \frac{(\dot{s}_2^{(kr)})^2}{k} = \frac{c^3 (\gamma_b - \gamma_a)^2}{\bar{k}} (\dot{s}_2^{(kr)})^2,
\end{aligned} \tag{3.7}$$

where

$$\dot{s}_1^{(kr)} = \dot{S}(\Delta\tilde{\gamma}), \quad \dot{s}_{con}^{(kr)} = \dot{S}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2), \quad \dot{s}_2^{(kr)} = \dot{S}(\Delta\tilde{\gamma}_2), \tag{3.8}$$

and $\dot{S}(\Delta\tilde{\gamma})$ is defined for all real $\Delta\tilde{\gamma}$ as the unique root, within the interval $-1 < \dot{S} = \dot{S}(\Delta\tilde{\gamma}) < 1$, to the equation

$$\dot{S}^3 - \bar{k}\Delta\tilde{\gamma}\dot{S}^2 - (1 + \frac{\bar{k}}{2})\dot{S} + \bar{k}\Delta\tilde{\gamma} = 0, \quad ((\Delta\tilde{\gamma} = 0) \Rightarrow (\dot{S} = 0)). \tag{3.9}$$

From (3.5) and (3.8) one also obtains

$$\dot{S}(\Delta\tilde{\gamma}) \begin{cases} > 0, & \text{when } \Delta\tilde{\gamma} > 0, \\ = 0, & \text{when } \Delta\tilde{\gamma} = 0, \\ < 0, & \text{when } \Delta\tilde{\gamma} < 0. \end{cases} \tag{3.10}$$

At the same time one finds that the first derivatives and second derivatives of $\dot{S}(\Delta\tilde{\gamma})$ obey

$$\dot{S}'(\Delta\tilde{\gamma}) \equiv \frac{d}{d\Delta\tilde{\gamma}} \dot{S}(\Delta\tilde{\gamma}) > 0, \quad \text{for all } \Delta\tilde{\gamma}, \tag{3.11}$$

and

$$\dot{S}''(\Delta\tilde{\gamma}) \equiv \frac{d^2}{d(\Delta\tilde{\gamma})^2} \dot{S}(\Delta\tilde{\gamma}) \begin{cases} < 0, & \text{when } \Delta\tilde{\gamma} > 0, \\ = 0, & \text{when } \Delta\tilde{\gamma} = 0, \\ > 0, & \text{when } \Delta\tilde{\gamma} < 0. \end{cases} \tag{3.12}$$

For small $\Delta\tilde{\gamma}$ one obtains from (3.9) that

$$\dot{S}(\Delta\tilde{\gamma}) = \left(\frac{2\bar{k}}{2+\bar{k}} \right) \Delta\tilde{\gamma} + O((\Delta\tilde{\gamma})^2). \tag{3.13}$$

In contrast, for small $\Delta\tilde{\gamma}$ under the maximum dissipation criterion, one finds from (3.8) of [1992L] that $\dot{S}(\Delta\tilde{\gamma}) = \Delta\tilde{\gamma} + O((\Delta\tilde{\gamma})^2)$, which coincides with (3.13) in the event that $\bar{k} = 2$. This indicates that, for infinitesimal pulses, the dynamical theory based upon the maximally dissipative solution criterion should be quantitatively similar to the dynamical theory based on the linear kinetic relation (3.4) with $\bar{k} = 2$.

4. Linear Kinetic Relation Solutions for the Concurrent Pulse in General

The (rr) , (lr) , and (ll) -cases can be treated in a similar fashion. In all cases, formulae (3.7)₂ and (3.8)₂ hold during the genuinely concurrent part of the encounter. If and when a portion of the encounter only involves the right moving pulse with strain increment $\Delta\gamma_1$ then (3.7)₁ and (3.8)₁ hold, whereas (3.7)₃ and (3.8)₃ govern those portions of any encounters that involve only the left moving pulse with strain increment $\Delta\gamma_2$. In order to determine which of the four possible cases is that which occurs, let

$$q = \frac{h}{2} - s_o. \quad (4.1)$$

Then one finds that the four cases occur according to

$$\begin{aligned} (rl): q > 0, & \quad ct_a(c + \dot{s}_1^{(kr)})(c - \dot{s}_{con}^{(kr)}) - ct_b(c + \dot{s}_1^{(kr)})(c + \dot{s}_{con}^{(kr)}) + 2q(c - \dot{s}_1^{(kr)})(c + \dot{s}_{con}^{(kr)}) > 0, \\ (rr): q > 0, & \quad ct_a(c + \dot{s}_1^{(kr)})(c - \dot{s}_{con}^{(kr)}) - ct_b(c + \dot{s}_1^{(kr)})(c + \dot{s}_{con}^{(kr)}) + 2q(c - \dot{s}_1^{(kr)})(c + \dot{s}_{con}^{(kr)}) < 0, \\ (lr): q < 0, & \quad ct_a(c - \dot{s}_2^{(kr)})(c - \dot{s}_{con}^{(kr)}) - ct_b(c - \dot{s}_2^{(kr)})(c + \dot{s}_{con}^{(kr)}) + 2q(c + \dot{s}_2^{(kr)})(c - \dot{s}_{con}^{(kr)}) < 0, \\ (ll): q < 0, & \quad ct_a(c - \dot{s}_2^{(kr)})(c - \dot{s}_{con}^{(kr)}) - ct_b(c - \dot{s}_2^{(kr)})(c + \dot{s}_{con}^{(kr)}) + 2q(c + \dot{s}_2^{(kr)})(c - \dot{s}_{con}^{(kr)}) > 0. \end{aligned} \quad (4.2)$$

5. Energy Loss for the Linear Kinetic Relation Solution in the (rl) -case

In this section we begin the examination of the question raised in the Introduction. For the (rl) case discussed in Sections 2 and 3, let $t_1^{(kr)}$, $t_{con}^{(kr)}$, and $t_2^{(kr)}$ denote the time duration of the encounters associated with the interaction periods Π_1 , Π_{con} , and Π_2 . These quantities are given in terms of $\dot{s}_1^{(kr)}$, $\dot{s}_{con}^{(kr)}$, and $\dot{s}_2^{(kr)}$ as

$$\begin{aligned} t_1^{(kr)} &= \frac{2q}{c + \dot{s}_1^{(kr)}}, \quad t_{con}^{(kr)} = \frac{ct_b(c + \dot{s}_1^{(kr)}) - 2q(c - \dot{s}_1^{(kr)})}{(c + \dot{s}_1^{(kr)})(c - \dot{s}_{con}^{(kr)})}, \\ t_2^{(kr)} &= \frac{ct_a(c + \dot{s}_1^{(kr)})(c - \dot{s}_{con}^{(kr)}) - ct_b(c + \dot{s}_1^{(kr)})(c + \dot{s}_{con}^{(kr)}) + 2q(c - \dot{s}_1^{(kr)})(c + \dot{s}_{con}^{(kr)})}{(c + \dot{s}_1^{(kr)})(c - \dot{s}_{con}^{(kr)})(c + \dot{s}_2^{(kr)})}, \end{aligned} \quad (5.1)$$

so that the total energy loss for the complete encounter process governed by (L.K.R.) is

$$\Delta E^{(kr)} = D_1^{(kr)} t_1^{(kr)} + D_{con}^{(kr)} t_{con}^{(kr)} + D_2^{(kr)} t_2^{(kr)}. \quad (5.2)$$

We now turn to consider the energy loss which would accompany two subsidiary problems.

The first problem is that in which only the right moving pulse, associated with strain increment $\Delta\gamma_1$, impinges upon the phase boundary. The encounter dynamics are again assumed to be governed by (L.K.R.). The phase boundary velocity and dissipation rate for this problem can simply be found by setting $\Delta\gamma_2=0$ in the previous development; consequently they are given by $\dot{s}_1^{(kr)}$ and $D_1^{(kr)}$. Similarly, the second problem is that in which only the left moving pulse, associated with strain increment $\Delta\gamma_2$, impinges upon the phase boundary with encounter dynamics governed

by (L.K.R.). Hence the phase boundary velocity and dissipation rate in this problem are given by $\dot{s}_2^{(kr)}$ and $D_2^{(kr)}$. It is, however, important to note that the time duration of the encounters are *not* given by $t_1^{(kr)}$ and $t_2^{(kr)}$, but rather are each of a longer duration due to the additional interaction time which was taken by the concurrent pulse in the original problem. These additional interaction times will be denoted respectively by $\delta t_1^{(kr)}$ and $\delta t_2^{(kr)}$; they are given by (see Figure 3):

$$\begin{aligned}\delta t_1^{(kr)} &= \frac{ct_b(c + \dot{s}_1^{(kr)}) - 2q(c - \dot{s}_1^{(kr)})}{(c + \dot{s}_1^{(kr)})(c - \dot{s}_1^{(kr)})}, \\ \delta t_2^{(kr)} &= \frac{ct_b(c + \dot{s}_1^{(kr)})(c + \dot{s}_{con}^{(kr)}) - 2q(c - \dot{s}_1^{(kr)})(c + \dot{s}_{con}^{(kr)})}{(c + \dot{s}_1^{(kr)})(c - \dot{s}_{con}^{(kr)})(c + \dot{s}_2^{(kr)})},\end{aligned}\quad (5.3)$$

so that the total energy losses in the two subsidiary problems are given by

$$\Delta E_{\Delta\gamma_1 \text{ only}}^{(kr)} = D_1^{(kr)}(t_1^{(kr)} + \delta t_1^{(kr)}), \quad \Delta E_{\Delta\gamma_2 \text{ only}}^{(kr)} = D_2^{(kr)}(t_2^{(kr)} + \delta t_2^{(kr)}). \quad (5.4)$$

Consequently, the difference in the energy loss between the original problem and the combined energy loss for the two subsidiary problems, is given by³

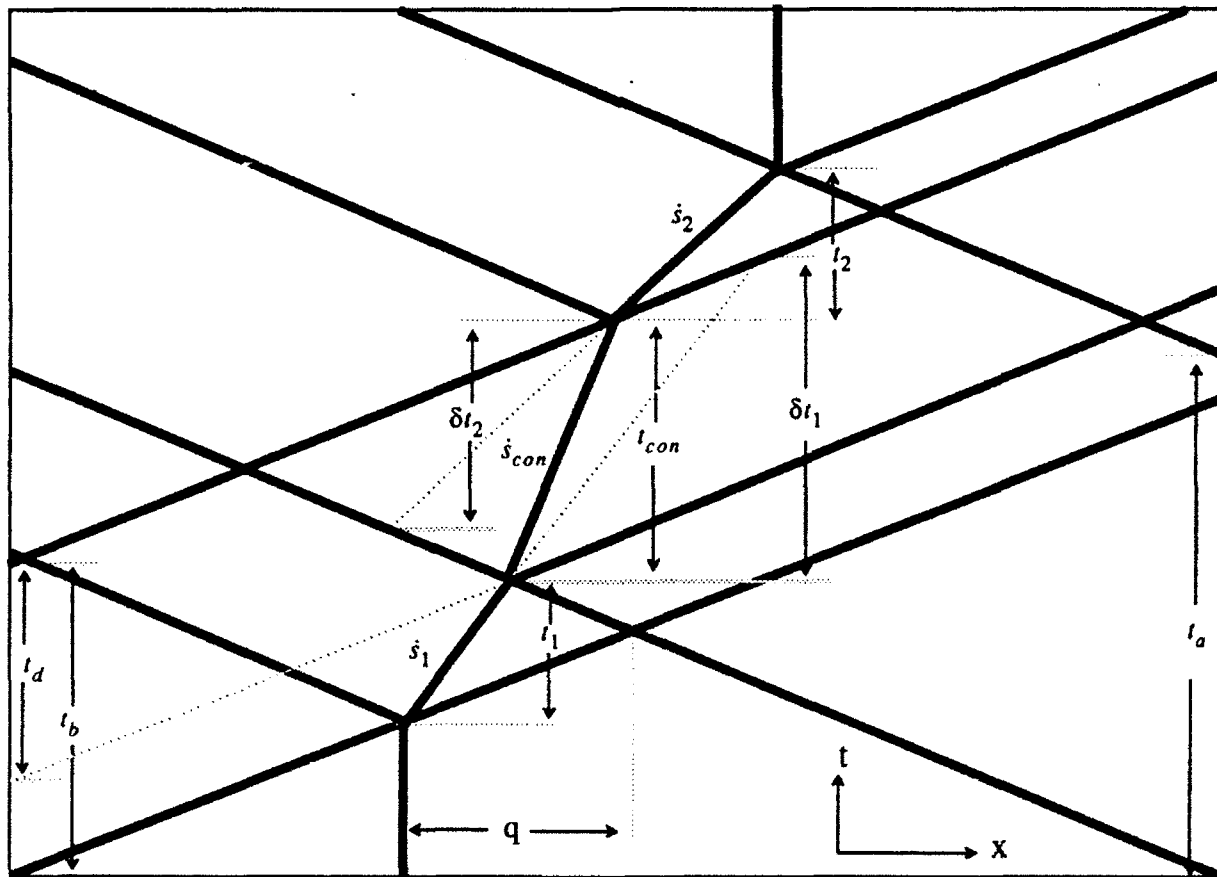


Fig. 3. The (rl) concurrent encounter.

$$\Upsilon_{(rl)} = \Delta E^{(kr)} - (\Delta E_{\Delta\gamma_1 \text{ only}}^{(kr)} + \Delta E_{\Delta\gamma_2 \text{ only}}^{(kr)}) = D_{con}^{(kr)} t_{con}^{(kr)} - D_1^{(kr)} \delta t_1^{(kr)} - D_2^{(kr)} \delta t_2^{(kr)}. \quad (5.5)$$

In order to develop a simple expression for $\Upsilon_{(rl)}$ it is convenient to introduce

$$t_d = t_b - \frac{2q(c - \dot{s}_1^{(kr)})}{c(c + \dot{s}_1^{(kr)})} > 0, \quad (5.6)$$

where $t_d > 0$ follows either from $t_{con}^{(kr)} > 0$ or else from its interpretation as a 'projected time' given in Figure 3. The interaction times $t_{con}^{(kr)}$, $\delta t_1^{(kr)}$ and $\delta t_2^{(kr)}$ are then given by

$$t_{con}^{(kr)} = \frac{1}{1 - s_{con}} t_d, \quad \delta t_1^{(kr)} = \frac{1}{1 - s_1} t_d, \quad \delta t_2^{(kr)} = \frac{(1 + s_{con})}{(1 - s_{con})(1 + s_2)} t_d. \quad (5.7)$$

Substituting from (5.7) into (5.5) and using (3.7) yields

$$\Upsilon_{(rl)} = \frac{c^3 t_d (\gamma_b - \gamma_a)^2}{\tilde{k}} \Phi(\dot{s}_1^{(kr)}, s_{con}^{(kr)}, \dot{s}_2^{(kr)}), \quad (5.8)$$

where

$$\Phi(\dot{s}_1^{(kr)}, s_{con}^{(kr)}, \dot{s}_2^{(kr)}) = \frac{\dot{s}_{con}^{(kr)2}}{(1 - s_{con})} - \frac{\dot{s}_1^{(kr)2}}{(1 - s_1)} - \frac{\dot{s}_2^{(kr)2}}{(1 + s_2)} \frac{\dot{s}_{con}^{(kr)}}{(1 - s_{con})}. \quad (5.9)$$

In view of (3.8) we define

$$\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) \equiv \Phi(\dot{S}(\Delta\tilde{\gamma}_1), \dot{S}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2), \dot{S}(\Delta\tilde{\gamma}_2)) = \Phi(\dot{s}_1^{(kr)}, s_{con}^{(kr)}, \dot{s}_2^{(kr)}). \quad (5.10)$$

Thus the question posed in the Introduction reduces, in the (rl) -case, to a determination of the sign of $\hat{\Phi}$. We note that if either $\Delta\tilde{\gamma}_1 = 0$ or $\Delta\tilde{\gamma}_2 = 0$ then the corresponding subsidiary problem is trivial and the remaining subsidiary problem is identical to the concurrent pulse problem, thus

$$\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) = 0, \quad \text{if } \Delta\tilde{\gamma}_1 = 0 \text{ or if } \Delta\tilde{\gamma}_2 = 0. \quad (5.11)$$

Analytically, the $\Delta\tilde{\gamma}_1 = 0$ case of (5.11) follows from (5.9) since $\Delta\tilde{\gamma}_1 = 0$ implies $\dot{s}_1^{(kr)} = 0$, $\dot{s}_{con}^{(kr)} = \dot{s}_2^{(kr)}$; while a similar argument gives the $\Delta\tilde{\gamma}_2 = 0$ case of (5.11).

We now begin an examination of the sign of $\hat{\Phi}$ in each of the four quadrants of the $(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ -plane. The following result establishes that this sign is negative in the second and fourth quadrants:

3. Note that $\Upsilon_{(rl)} = \Upsilon_{(rl)}^{(kr)}$ which is in general different from $\Upsilon_{(rl)} = \Upsilon_{(rl)}^{(md)}$ appearing in (5.5) of [1992L]. For legibility, the (kr) superscript is implied in what follows. A similar remark applies to subsequent functions, such as Φ and $\hat{\Phi}$ appearing in (5.8) and (5.10) respectively. We note in passing that the functional dependence upon k and \tilde{k} has not always been acknowledged in the argument lists, eg. $S(\Delta\tilde{\gamma}) = S(\Delta\tilde{\gamma}, \tilde{k})$.

Theorem 1. If either $\Delta\tilde{\gamma}_1 < 0, \Delta\tilde{\gamma}_2 > 0$ or $\Delta\tilde{\gamma}_1 > 0, \Delta\tilde{\gamma}_2 < 0$, then $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) < 0$.

Proof. Consider separately the five cases corresponding to the four octants which comprise the second and fourth quadrants, along with the diagonal line separating these octants.

Case 1: the diagonal line. Here $\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2 = 0$, yielding $\frac{z(kr)}{s_{con}} = 0$ so that the first term of the right hand side of (5.9) is zero. The second term and the third term of the right hand side of (5.9) then give that $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) < 0$.

Case 2: the N-NW octant. Here $\Delta\tilde{\gamma}_1 < 0, \Delta\tilde{\gamma}_2 > 0$ and $\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2 > 0$, in which case $\frac{z(kr)}{s_1} < 0 < \frac{z(kr)}{s_{con}} < \frac{z(kr)}{s_2}$. The first term and the third term of the right hand side of (5.9) then combine to give

$$\begin{aligned} & \frac{\frac{z(kr)^2}{s_{con}}}{(1 - \frac{z(kr)}{s_{con}})} - \frac{\frac{z(kr)^2}{s_2} \frac{z(kr)}{(1 + \frac{z(kr)}{s_{con}})}}{(1 + \frac{z(kr)}{s_2}) (1 - \frac{z(kr)}{s_{con}})} \\ &= \frac{1}{(1 + \frac{z(kr)}{s_2}) (1 - \frac{z(kr)}{s_{con}})} \left(\frac{z(kr)^2}{s_{con}} \frac{z(kr)}{(1 + \frac{z(kr)}{s_2})} - \frac{z(kr)^2}{s_2} \frac{z(kr)}{(1 + \frac{z(kr)}{s_{con}})} \right) \\ &= \frac{1}{(1 + \frac{z(kr)}{s_2}) (1 - \frac{z(kr)}{s_{con}})} \left(\frac{z(kr)^2}{s_{con}} - \frac{z(kr)^2}{s_2} + \frac{z(kr)}{s_2} \frac{z(kr)}{s_{con}} \frac{z(kr)}{(s_{con} - s_2)} \right) < 0. \end{aligned} \quad (5.12)$$

Since the second term of the right hand side of (5.9) is negative, the result (5.12) then gives $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) < 0$.

Case 3: the W-NW octant. Here $\Delta\tilde{\gamma}_1 < 0, \Delta\tilde{\gamma}_2 > 0$ and $\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2 < 0$, in which case $\frac{z(kr)}{s_1} < \frac{z(kr)}{s_{con}} < 0 < \frac{z(kr)}{s_2}$. The first term and the second term of the right hand side of (5.9) then combine to give

$$\begin{aligned} & \frac{\frac{z(kr)^2}{s_{con}}}{(1 - \frac{z(kr)}{s_{con}})} - \frac{\frac{z(kr)^2}{s_1}}{(1 - \frac{z(kr)}{s_1})} \\ &= \frac{1}{(1 - \frac{z(kr)}{s_1}) (1 - \frac{z(kr)}{s_{con}})} \left(\frac{z(kr)^2}{s_{con}} \frac{z(kr)}{(1 - \frac{z(kr)}{s_1})} - \frac{z(kr)^2}{s_1} \frac{z(kr)}{(1 - \frac{z(kr)}{s_{con}})} \right) \\ &= \frac{1}{(1 - \frac{z(kr)}{s_1}) (1 - \frac{z(kr)}{s_{con}})} \left(\frac{z(kr)^2}{s_{con}} - \frac{z(kr)^2}{s_1} + \frac{z(kr)}{s_1} \frac{z(kr)}{s_{con}} \frac{z(kr)}{(s_1 - s_{con})} \right) < 0. \end{aligned} \quad (5.13)$$

Since the third term of the right hand side of (5.9) is negative, the result (5.13) then gives $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) < 0$.

Cases 4 and 5 corresponding to the E-SE and S-SE octants can be treated similarly.

QED

Theorem 1 indicates for the (*rl*) case that the concurrent pulse encounter suffers the lesser energy loss in the event that the incoming pulses are of opposite signs. As mentioned in the Introduction, a similar result was obtained in [1992L] for the case in which the phase boundary motion is governed by (M.D.C.) instead of the kinetic criterion (3.4) considered here.

Before examining the first and third quadrants of the $(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ -plane, we shall briefly consider the behavior of $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ near the origin. Let partial derivatives of $\hat{\Phi}$ be denoted by numerical subscripts, e.g. $\hat{\Phi}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) = \frac{\partial}{\partial \Delta\tilde{\gamma}_1} \hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$, then (5.11) gives $\hat{\Phi}(0,0) = 0$, $\hat{\Phi}_1(0,0) = 0$, $\hat{\Phi}_2(0,0) = 0$, $\hat{\Phi}_{11}(0,0) = 0$, $\hat{\Phi}_{22}(0,0) = 0$, while (3.8), (3.9), (5.9), (5.10) gives $\hat{\Phi}_{12}(0,0) = \frac{2}{(1/\tilde{k} + 1/2)^2}$, so that

$$\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) = \frac{2\Delta\tilde{\gamma}_1\Delta\tilde{\gamma}_2}{(1/\tilde{k} + 1/2)^2} + O((\Delta\tilde{\gamma})^3) \quad (5.14)$$

yielding saddle point behavior near the origin. This result corroborates Theorem 1 and also indicates that $\hat{\Phi} > 0$ in the first and third quadrants provided that $(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ is sufficiently close to $(0,0)$. Again, similar qualitative results were obtained in [1992L] for the (M.D.C.)-case. Also, as anticipated, the special case of $\tilde{k} = 2$ in (5.14) gives the same local result as that for the maximally dissipative problem (*viz* (5.11) of [1992L]).

The qualitative behavior of solutions to the problem under study here departs from the behavior of the solutions governed by (M.D.C.) as studied in [1992L] for incoming pulse incre-

Table 1. Values of $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ for $\tilde{k} = 1$.

$\Delta\tilde{\gamma}_2 \backslash \Delta\tilde{\gamma}_1$	-1.000	-0.800	-0.600	-0.400	-0.200	0.000	0.200	0.400	0.600	0.800	1.000
1.000	-0.418	-0.410	-0.368	-0.286	-0.162	0.000	0.187	0.378	0.545	0.665	0.717
0.800	-0.314	-0.316	-0.290	-0.231	-0.135	0.000	0.168	0.353	0.531	0.677	0.770
0.600	-0.216	-0.222	-0.208	-0.169	-0.101	0.000	0.134	0.294	0.463	0.617	0.735
0.400	-0.129	-0.136	-0.130	-0.108	-0.065	0.000	0.091	0.208	0.340	0.474	0.591
0.200	-0.056	-0.061	-0.059	-0.050	-0.031	0.000	0.045	0.105	0.178	0.259	0.336
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
-0.200	0.040	0.045	0.047	0.041	0.026	0.000	-0.041	-0.098	-0.174	-0.268	-0.374
-0.400	0.066	0.076	0.080	0.072	0.047	0.000	-0.075	-0.183	-0.330	-0.518	-0.742
-0.600	0.078	0.092	0.100	0.092	0.062	0.000	-0.102	-0.253	-0.462	-0.735	-1.072
-0.800	0.081	0.097	0.107	0.102	0.070	0.000	-0.121	-0.304	-0.563	-0.908	-1.342
-1.000	0.075	0.093	0.106	0.103	0.072	0.000	-0.130	-0.334	-0.629	-1.028	-1.539

Table 2. Values of $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ for $\tilde{k}=10$.

$\Delta\tilde{\gamma}_2 \backslash \Delta\tilde{\gamma}_1$	1.000	-0.800	-0.600	-0.400	-0.200	0.000	0.200	0.400	0.600	0.800	1.000
1.000	-0.664	-0.787	-0.738	-0.552	-0.277	0.000	0.089	-0.081	-0.414	-0.830	-1.288
0.800	-0.418	-0.595	-0.673	-0.558	-0.295	0.000	0.122	-0.007	-0.294	-0.661	-1.069
0.600	-0.237	-0.357	-0.496	-0.506	-0.301	0.000	0.161	0.088	-0.134	-0.432	-0.768
0.400	-0.122	-0.179	-0.268	-0.348	-0.265	0.000	0.197	0.197	0.066	-0.130	-0.359
0.200	-0.046	-0.065	-0.097	-0.140	-0.144	0.000	0.182	0.252	0.233	0.170	0.085
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
-0.200	0.010	0.016	0.027	0.044	0.059	0.000	-0.272	-0.662	-0.999	-1.277	-1.525
-0.400	-0.008	-0.001	0.012	0.033	0.056	0.000	-0.351	-1.084	-2.005	-2.927	-3.821
-0.600	-0.034	-0.028	-0.015	0.009	0.039	0.000	-0.321	-1.063	-2.238	-3.757	-5.433
-0.800	-0.060	-0.054	-0.041	-0.014	0.023	0.000	-0.283	-0.941	-1.993	-3.567	-5.722
-1.000	-0.084	-0.078	-0.063	-0.034	0.009	0.000	-0.254	-0.847	-1.747	-3.035	-4.990

ments $(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ which lie away from the origin in the first and third quadrants. We have numerically calculated $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ for various pairs $(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ and various mobility values \tilde{k} . Some of these results are displayed in Table 1 ($\tilde{k} = 1$) and Table 2 ($\tilde{k} = 10$). It is to be noted for the $\tilde{k} = 1$ case of Table 1 that $\hat{\Phi} > 0$ in the first and third quadrants for all of the tabulated pairs $(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$. In contrast, for the $\tilde{k} = 10$ case of Table 2 it is found that $\hat{\Phi} < 0$ in the first and third quadrants for certain pairs $(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$. Similar data for the cases $\tilde{k} = 2$ and 100 indicates that the extent of the $(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ -domain in the first and third quadrant on which $\hat{\Phi} < 0$ increases with \tilde{k} . The following Theorem provides a certain characterization of the domain in the first and third quadrants on which $\hat{\Phi} > 0$ in terms of the normalized phase boundary speed $\dot{\tilde{s}}_{con}^{(kr)} = \dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2)$.

Theorem 2. If either $\Delta\tilde{\gamma}_1 > 0, \Delta\tilde{\gamma}_2 > 0$ or $\Delta\tilde{\gamma}_1 < 0, \Delta\tilde{\gamma}_2 < 0$, such that $|\dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2)| < 0.5$, then

$$\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) > 0.$$

Proof. Using (3.5) to rearrange (5.9), one finds that (5.10) can be expressed as

$$\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) = \hat{X}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2, \tilde{k}) + \hat{Y}(\Delta\tilde{\gamma}_1)\hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) + \hat{Y}(\Delta\tilde{\gamma}_2)\hat{Z}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2)\hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2). \quad (5.15)$$

where

$$\hat{X}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2, \tilde{k}) = X(\dot{\tilde{S}}(\Delta\tilde{\gamma}_1), \dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2), \dot{\tilde{S}}(\Delta\tilde{\gamma}_2), \tilde{k}),$$

$$\hat{Y}(\Delta\tilde{\gamma}_1) = Y(\dot{\tilde{S}}(\Delta\tilde{\gamma}_1)), \quad \hat{Z}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2) = Z(\dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2)), \quad (5.16)$$

$$\hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) = W_1(\dot{\tilde{S}}(\Delta\tilde{\gamma}_1), \dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2)), \quad \hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) = W_2(\dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2), \dot{\tilde{S}}(\Delta\tilde{\gamma}_2)).$$

with

$$\begin{aligned}
 X(s_1^{(kr)}, s_{con}^{(kr)}, s_2^{(kr)}, \bar{k}) &= \frac{2(1+s_{con}^{(kr)})(s_{con}^{(kr)})(s_1^{(kr)} + s_2^{(kr)} - s_{con}^{(kr)})}{\bar{k}}, \\
 Y(s_1^{(kr)}) &= \frac{s_1^{(kr)}}{(1-s_1^{(kr)})^2}, \quad Z(s_{con}^{(kr)}) = \frac{1+s_{con}^{(kr)}}{1-s_{con}^{(kr)}}, \\
 W_1(s_1^{(kr)}, s_{con}^{(kr)}) &= s_{con}^{(kr)}(1+s_{con}^{(kr)}) - s_1^{(kr)}(1+s_1^{(kr)}), \\
 W_2(s_{con}^{(kr)}, s_2^{(kr)}) &= s_{con}^{(kr)}(1-s_{con}^{(kr)}) - s_2^{(kr)}(1-s_2^{(kr)}).
 \end{aligned} \tag{5.17}$$

We now examine the sign of each of the five functions appearing in (5.17). The following lemma addresses the function $\hat{X}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2, \bar{k})$.

Lemma A. $\hat{X}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2, \bar{k}) > 0$, if $\Delta\tilde{\gamma}_1 \Delta\tilde{\gamma}_2 > 0$.

Proof of Lemma A. Since $\bar{k} > 0$ and $1 + s_{con}^{(kr)} > 0$ it follows from (5.17)₁ that

$$\hat{X}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2, \bar{k}) > 0, \quad \text{if } (s_{con}^{(kr)})(s_1^{(kr)} + s_2^{(kr)} - s_{con}^{(kr)}) > 0. \tag{5.18}$$

In the first quadrant, $\Delta\tilde{\gamma}_1 > 0$ and $\Delta\tilde{\gamma}_2 > 0$, one has $s_{con}^{(kr)} > 0$, and we examine the remaining factor $s_1^{(kr)} + s_2^{(kr)} - s_{con}^{(kr)}$ of (5.18) by considering separately the two cases: $\Delta\tilde{\gamma}_1 \geq \Delta\tilde{\gamma}_2$ and $\Delta\tilde{\gamma}_1 < \Delta\tilde{\gamma}_2$. In the first case, $\Delta\tilde{\gamma}_1 \geq \Delta\tilde{\gamma}_2 > 0$, an application of the Mean Value Theorem gives

$$\begin{aligned}
 s_{con}^{(kr)} &= \tilde{S}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2) = \tilde{S}(\Delta\tilde{\gamma}_1) + \Delta\tilde{\gamma}_2 \tilde{S}'(a), \quad \text{with } 0 < \Delta\tilde{\gamma}_1 < a < \Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2, \\
 s_2^{(kr)} &= \tilde{S}(\Delta\tilde{\gamma}_2) = \tilde{S}(0) + \Delta\tilde{\gamma}_2 \tilde{S}'(b) = \Delta\tilde{\gamma}_2 \tilde{S}'(b), \quad \text{with } 0 < b < \Delta\tilde{\gamma}_2,
 \end{aligned} \tag{5.19}$$

so that

$$s_1^{(kr)} + s_2^{(kr)} - s_{con}^{(kr)} = \Delta\tilde{\gamma}_2 (\tilde{S}'(b) - \tilde{S}'(a)), \quad \text{with } 0 < b < \Delta\tilde{\gamma}_2 \leq \Delta\tilde{\gamma}_1 < a < \Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2. \tag{5.20}$$

Now (3.11) and (3.12) give that $a > b > 0 \Rightarrow \tilde{S}'(b) > \tilde{S}'(a) > 0$ wherein (5.20) yields $s_1^{(kr)} + s_2^{(kr)} - s_{con}^{(kr)} > 0$ which establishes the required condition in (5.18) for the case $\Delta\tilde{\gamma}_1 \geq \Delta\tilde{\gamma}_2 > 0$. The required condition in (5.18) can be established in a similar fashion for the other first quadrant case $\Delta\tilde{\gamma}_2 > \Delta\tilde{\gamma}_1 > 0$. The case for the third quadrant follows in a similar fashion.

QED Lemma A

Now since $-1 < s_1^{(kr)} < 1$ with $\text{sign}(s_1^{(kr)}) = \text{sign}(\Delta\tilde{\gamma}_1)$ it follows from (5.17)₂ that

$$\text{sign}(\hat{Y}(\Delta\tilde{\gamma}_1)) = \text{sign}(\Delta\tilde{\gamma}_1) \tag{5.21}$$

while (5.17)₃ with $-1 < s_{con}^{(kr)} < 1$ gives

$$\hat{Z}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2) > 0. \tag{5.22}$$

The next lemma addresses the functions $\hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ and $\hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$.

Lemma B. If $\Delta\tilde{\gamma}_1 > 0, \Delta\tilde{\gamma}_2 > 0$ such that $|\dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2)| < 0.5$, then $\hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) > 0$ and $\hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) > 0$.

If $\Delta\tilde{\gamma}_1 < 0, \Delta\tilde{\gamma}_2 < 0$ such that $|\dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2)| < 0.5$, then $\hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) < 0$ and $\hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) < 0$.

Proof of Lemma B. We write

$$\begin{aligned}\hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) &= W_1(\dot{s}_1^{(kr)}, \dot{s}_{con}^{(kr)}) = G_1(\dot{s}_{con}^{(kr)}) - G_1(\dot{s}_1^{(kr)}), & G_1(\dot{s}) &\equiv \dot{s}(1 + \dot{s}), \\ \hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) &= W_2(\dot{s}_{con}^{(kr)}, \dot{s}_2^{(kr)}) = G_2(\dot{s}_{con}^{(kr)}) - G_2(\dot{s}_2^{(kr)}), & G_2(\dot{s}) &\equiv \dot{s}(1 - \dot{s}),\end{aligned}\quad (5.23)$$

and note that both $G_1(\dot{s})$ and $G_2(\dot{s})$ are monotonically increasing for $|\dot{s}| < 0.5$. Therefore if $\Delta\tilde{\gamma}_1 > 0, \Delta\tilde{\gamma}_2 > 0$ such that $|\dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2)| < 0.5$ it then follows that

$$\begin{aligned}\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2 > \Delta\tilde{\gamma}_1 > 0 &\Rightarrow 0.5 > \dot{s}_{con}^{(kr)} > \dot{s}_1^{(kr)} > 0 \Rightarrow G_1(\dot{s}_{con}^{(kr)}) > G_1(\dot{s}_1^{(kr)}) \Rightarrow \hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) > 0, \\ \Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2 > \Delta\tilde{\gamma}_2 > 0 &\Rightarrow 0.5 > \dot{s}_{con}^{(kr)} > \dot{s}_2^{(kr)} > 0 \Rightarrow G_2(\dot{s}_{con}^{(kr)}) > G_2(\dot{s}_2^{(kr)}) \Rightarrow \hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) > 0.\end{aligned}\quad (5.24)$$

Similarly if $\Delta\tilde{\gamma}_1 < 0, \Delta\tilde{\gamma}_2 < 0$ such that $|\dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2)| < 0.5$ it follows that

$$\begin{aligned}\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2 < \Delta\tilde{\gamma}_1 < 0 &\Rightarrow -0.5 < \dot{s}_{con}^{(kr)} < \dot{s}_1^{(kr)} < 0 \Rightarrow G_1(\dot{s}_{con}^{(kr)}) < G_1(\dot{s}_1^{(kr)}) \Rightarrow \hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) < 0, \\ \Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2 < \Delta\tilde{\gamma}_2 < 0 &\Rightarrow -0.5 < \dot{s}_{con}^{(kr)} < \dot{s}_2^{(kr)} < 0 \Rightarrow G_2(\dot{s}_{con}^{(kr)}) < G_2(\dot{s}_2^{(kr)}) \Rightarrow \hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) < 0.\end{aligned}\quad (5.25)$$

QED Lemma B

Therefore (5.21), (5.22) and Lemma B give, under the hypotheses of the Theorem, that

$$\hat{Y}(\Delta\tilde{\gamma}_1)\hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) > 0, \quad \hat{Y}(\Delta\tilde{\gamma}_2)\hat{Z}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2)\hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) > 0. \quad (5.26)$$

This result, in conjunction with Lemma A and (5.15), then establishes the Theorem.

QED

In Table 3 we display the phase boundary speed $\dot{s}_{con}^{(kr)}$ corresponding to the $\bar{k} = 10$ case for which Φ was previously given in Table 2. Examination of the bold entries of these two tables indicates numerical consistency with Theorem 2. In addition, it is also clear that the converse to Theorem 2 does not hold.

6. Energy Loss for the General Concurrent Pulse Problem Governed by the Linear Kinetic Relation

The energy losses for the (rr) , (lr) , and (ll) -cases can be determined in a corresponding way. The energy loss differences analogous to $\Upsilon_{(rl)}$ for these other cases are found to be given by

Table 3. Values of $\tilde{s}_{con}^{(kr)}$ for $\tilde{k}=10$.

$\Delta\tilde{\gamma}_2 \backslash \Delta\tilde{\gamma}_1$	-1.000	-0.800	-0.600	-0.400	-0.200	0.000	0.200	0.400	0.600	0.800	1.000
1.000	0.000	0.307	0.513	0.637	0.714	0.765	0.801	0.828	0.849	0.865	0.878
0.800	-0.307	0.000	0.307	0.513	0.637	0.714	0.765	0.801	0.828	0.849	0.865
0.600	-0.513	-0.307	0.000	0.307	0.513	0.637	0.714	0.765	0.801	0.828	0.849
0.400	-0.637	-0.513	-0.307	0.000	0.307	0.513	0.637	0.714	0.765	0.801	0.828
0.200	-0.714	-0.637	-0.513	-0.307	0.000	0.307	0.513	0.637	0.714	0.765	0.801
0.000	-0.765	-0.714	-0.637	-0.513	-0.307	0.000	0.307	0.513	0.637	0.714	0.765
-0.200	-0.801	-0.765	-0.714	-0.637	-0.513	-0.307	0.000	0.307	0.513	0.637	0.714
-0.400	-0.828	-0.801	-0.765	-0.714	-0.637	-0.513	-0.307	0.000	0.307	0.513	0.637
-0.600	-0.849	-0.828	-0.801	-0.765	-0.714	-0.637	-0.513	-0.307	0.000	0.307	0.513
-0.800	-0.865	-0.849	-0.828	-0.801	-0.765	-0.714	-0.637	-0.513	-0.307	0.000	0.307
-1.000	-0.878	-0.865	-0.849	-0.828	-0.801	-0.765	-0.714	-0.637	-0.513	-0.307	0.000

$$\begin{aligned}
 (rr): \quad \Upsilon_{(rr)} &= \frac{c^3 t_a (\gamma_b - \gamma_a)^2}{\tilde{k}} \hat{\Psi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2), \\
 (lr): \quad \Upsilon_{(lr)} &= \frac{c^3 t_c (\gamma_b - \gamma_a)^2}{\tilde{k}} \hat{\Psi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2), \\
 (ll): \quad \Upsilon_{(ll)} &= \frac{c^3 t_b (\gamma_b - \gamma_a)^2}{\tilde{k}} \hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2),
 \end{aligned} \tag{6.1}$$

with

$$\hat{\Psi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) = \hat{\Phi}(-\Delta\tilde{\gamma}_2, -\Delta\tilde{\gamma}_1). \tag{6.2}$$

In the (lr) -case, another projected time $t_c \equiv t_a + \frac{2q(c + \dot{s}_2^{(kr)})}{c(c - \dot{s}_2^{(kr)})}$ has been introduced.

In particular, $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ and $\hat{\Psi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ are each negative in the second and fourth quadrants, whereas neither sign is excluded in the first and third quadrants. Thus we now summarize our findings for the concurrent pulse problem with phase boundary speed governed by the linear kinetic relation (3.4):

The concurrent pulse encounter suffers the lesser energy loss in the event that both incoming pulses are of opposite sign, whereas knowledge that the incoming pulses are of the same sign is not sufficient to conclude that the concurrent pulse encounter suffers the greater energy loss. However if the incoming pulses are of the same sign, and each is sufficiently small, then the concurrent pulse encounter suffers the greater energy loss.

Finally, we recall from Section 2 that we have not seriously considered the effect of the additional restrictions upon the phase boundary velocities which are necessary to ensure that the strain values in (2.5) remain in the segregation intervals associated with their respective phases. In particular, we have not explicitly constructed an example in which: the incoming pulses are of the same sign, the concurrent pulse suffers the lesser energy loss (implying by Theorem 2 that $|\dot{s}^{(kr)}_{con}| \geq 0.5$), and $\dot{s}_1^{(kr)} \in \mathfrak{S}_1$, $\dot{s}_{con}^{(kr)} \in \mathfrak{S}_{con}$, $\dot{s}_2^{(kr)} \in \mathfrak{S}_2$.

References

- [1973D] Dafermos, C.M., The entropy rate admissibility criterion for solutions of hyperbolic conservation laws, *J. Diff. Eqs.* **14**, 202-212.
- [1975E] Ericksen, J.L., Equilibrium of bars, *J. Elasticity* **5**, 191-201.
- [1980J] James, R.D., The propagation of phase boundaries in elastic bars, *Arch. Rational Mech. Anal.* **73**, 125-158.
- [1983H] Hagan, R. and M. Slemrod, The viscosity-capillarity criterion for shocks and phase transitions, *Arch. Rational Mech. Anal.* **83**, 333-361.
- [1986H] Hattori, H., The Riemann problem for a van der Waals fluid with entropy rate admissibility criterion-isothermal case, *Arch. Rational Mech. Anal.* **92**, 247-263.
- [1987T] Truskinovsky, L., Dynamics of nonequilibrium phase boundaries in a heat conducting nonlinear elastic medium, *J. Appl. Math. Mech. (PMM)* **51** (1987) 777-784.
- [1990G] Gurtin, M.E. and A. Struthers, Multiphase thermomechanics with interfacial structure 3. evolving phase boundaries in the presence of bulk deformation, *Arch. Rational Mech. Anal.* **112**, 97-160.
- [1991A] Abeyaratne, R. and J.K. Knowles, Kinetic relations and the propagation of phase boundaries in solids, *Arch. Rational Mech. Anal.* **114**, 119-154.
- [1991F] Fried, E., Stability of a two-phase process involving a planar phase boundary in an elastic solid, to appear in *J. Elasticity*.
- [1991L] Lin, J. and T.J. Pence, On the energy dissipation due to wave ringing in non-elliptic elastic materials, accepted for publication in *Journal of Nonlinear Science*.
- [1991P] Pence, T.J., On the encounter of an acoustic shear pulse with a phase boundary in an elastic material: reflection and transmission behavior, *J. Elasticity* **25**, 31-74.
- [1991PP] Pence, T.J., On the encounter of an acoustic shear pulse with a phase boundary in an elastic material: energy and dissipation, *J. Elasticity* **26**, 95-146.
- [1991T] Truskinovsky, L., Kinks vs. shocks, to appear in *Shock Induced Transitions and Phase Structures in General Media* (R. Fosdick, E. Dunn and M. Slemrod, eds.) Springer-Verlag.
- [1992L] Lin, J. and T.J. Pence, Energy dissipation in an elastic material containing a mobile phase boundary subjected to concurrent dynamic pulses, *Transactions of the Ninth Army Conference on Applied Mathematics and Computing*, 437-450.
- [1992P] Pence, T.J., On the mechanical dissipation of solutions to the Riemann problem for impact involving a two-phase elastic material, *Arch. Rational Mech. Anal.* **117**, 1-52.

A FUNCTION WHOSE VALUES ARE INTEGERS, II

JOSEPH ARKIN, DAVID C. ARNEY, and EDITH H. LUCHINS*
Department of Mathematical Sciences
United States Military Academy
West Point, NY 10996-1786

Abstract: In this paper we prove the following:

For any value of k , there exists a value of n , for which

(A)
$$\left(\frac{\sum_{r=1}^n r^{2k}}{\sum_{r=1}^n r^2} \right)$$

is an integer.

Introduction. In [1], using determinants of certain triangular matrices, the following was proved.

Theorem. For any positive integers k and n ,

(1)
$$\left(\frac{(2k+1)!}{3 \cdot 2^k k!} \right) \left(\frac{\sum_{r=1}^n r^{2k}}{\sum_{r=1}^n r^2} \right)$$

is an integer.

The proof is repeated here. For any integers k and w , $k > 0$, notice that

* Dr. Luchins spent the 1991-1992 academic year as the Distinguished Visiting Professor of Mathematics at the United States Military Academy and has since returned to Rensselaer Polytechnic Institute

$$(2) \quad (a) \quad (w+1)^{2k+1} - w^{2k+1} = 1 + \sum_{r=1}^{2k} \binom{2k+1}{r} w^{2k+1-r},$$

$$(b) \quad (w-1)^{2k+1} - w^{2k+1} = -1 + \sum_{r=1}^{2k} (-1)^r \binom{2k+1}{r} w^{2k+1-r},$$

where

$$\binom{p}{q} = \frac{p!}{q!(p-q)!} \quad (q = 0, 1, \dots, p).$$

For any positive integers m and n , let $S(m, n) = 1^m + 2^m + \dots + n^m$. If in (2a) we let $w = 0, 1, 2, \dots, n$, and add, we obtain

$$(3) \quad (n+1)^{2k+1} = \left(\sum_{r=1}^{2k} \binom{2k+1}{r} S(2k+1-r, n) \right) + n+1.$$

Just as we found (3) from (2a), from (2b) we get

$$(4) \quad -n^{2k+1} = \left(\sum_{r=1}^{2k} (-1)^r \binom{2k+1}{r} S(2k+1-r, n) \right) - n.$$

Replacing k by j and subtracting (4) from (3), we obtain

$$(5) \quad a_{2j+1} = 2 \sum_{r=1}^j \binom{2j+1}{2r-1} S(2j+2-2r, n),$$

where

$$(6) \quad a_{2j+1} = (n+1)^{2j+1} + n^{2j+1} - (2n+1).$$

Next, we consider the k equations obtained from (5) by successively replacing j by $1, 2, \dots, k$. With $l = 1, 2, \dots, k$, these k equations in the k unknowns $S(2l, n)$ can be solved by Cramer's rule to obtain

$$(7) \quad 2(D_1)S(2k, n) = (-1)^{k+1} D_2,$$

where D_1 and D_2 are the determinants given below:

$$D_1 = \begin{vmatrix} \binom{2k+1}{1} & \binom{2k+1}{3} & \cdots & \binom{2k+1}{2k-3} & \binom{2k+1}{2k-1} \\ 0 & \binom{2k-1}{1} & \cdots & \binom{2k-1}{2k-5} & \binom{2k-1}{2k-3} \\ 0 & 0 & \cdots & \binom{2k-3}{2k-7} & \binom{2k-3}{2k-5} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \binom{5}{1} & \binom{5}{3} \\ 0 & 0 & \cdots & 0 & \binom{3}{1} \end{vmatrix}$$

$$D_2 = \begin{vmatrix} a_{2k+1} & \binom{2k+1}{3} & \cdots & \binom{2k+1}{2k-3} & \binom{2k+1}{2k-1} \\ a_{2k-1} & \binom{2k-1}{1} & \cdots & \binom{2k-1}{2k-5} & \binom{2k-1}{2k-3} \\ a_{2k-3} & 0 & \cdots & \binom{2k-3}{2k-7} & \binom{2k-3}{2k-5} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_5 & 0 & \cdots & \binom{5}{1} & \binom{5}{3} \\ a_3 & 0 & \cdots & 0 & \binom{3}{1} \end{vmatrix}$$

Since D_1 is the determinant of a triangular matrix, it is simply the product of the diagonal entries. Hence

$$D_1 = \prod_{r=1}^k (2r+1) = (2k+1)! / 2^k k!.$$

Equation (7) then becomes

$$(8) \quad 6 \left(\frac{(2k+1)!}{3 \cdot 2^k \cdot k!} \right) S(2k, n) = D_2.$$

To complete the proof it suffices to show that D_2 is a multiple of $6S(2, n)$. This will be so if every entry in a column of D_2 is such a multiple.

To this end let $u = 6S(2, n) = n(n+1)(2n+1)$. Then we show that for any positive integer j

$$(9) \quad a_{2j+1} \equiv 0 \pmod{u},$$

which shows that u divides every entry in the first column of D_2 . To reduce the problem still further, notice that $n, n+1$, and $2n+1$ are relatively prime in pairs, so that the congruence (9) is equivalent to the three congruences $a_{2j+1} \equiv 0 \pmod{m}$, ($m = n, n+1, 2n+1$). Although our notation has hidden the fact that a_{2j+1} depends on n , equation (6) shows that these congruences become

$$\begin{aligned} (n+1)^{2k+1} &\equiv 1 \pmod{n}, \\ (n+1)^{2k+1} + n^{2k+1} &\equiv 0 \pmod{2n+1}, \\ n^{2k+1} &\equiv n \pmod{n+1}, \end{aligned}$$

which are easily verified. This completes the proof of (1).

Proof of (A). It is evident that the denominator in (A) equals $\frac{n(n+1)(2n+1)}{6} = D$. In (1), it is easy to show

$$\text{that } \frac{(2k+1)!}{3 \cdot 2^k k!} = 1, 5, 5, \dots, (2k+1) = J.$$

Let $n = (2k+1)! + 1$ so that $n+1 = 2 \left[\frac{(2k+1)!}{2+1} \right]$ and $2n+1 = 3 \left[\frac{2(2k+1)!}{3+1} \right]$. Now, it is evident that the only common factor of D and J is 1, which completes the proof of (A).

REFERENCES

- [1] Joseph Arkin, "A function whose values are integers", Mathematics Magazine, Vol. 38, No. 4, September 1965.

ACKNOWLEDGEMENTS

The authors wish to thank Colonel Frank R. Giordano, Head of the Department of Mathematical Sciences at the U.S. Military Academy, West Point, NY, for his interest and support of this paper.

REVERSE DIGIT CONSTRUCTIONS OF PERFECT, MAGIC, AND
DOUBLY MAGIC CUBES

Joseph Arkin
David C. Arney
Frank R. Giordano
Rickey A. Kolb
United States Military Academy
West Point, NY 10996-1786
and
Paul Smith
University of British Columbia
Vancouver, British Columbia V6T 1W5

ABSTRACT

In this paper we introduce the concept of reverse digit pairs. We then exhibit an orthogonal, perfect, magic and doubly magic 4-cube of order 8, with the property that each row contains reverse digit pairs. In addition, orthogonal cubes with this reverse digit property are discussed and in particular an orthogonal, reversed digit 4-cube of order 10 is presented.

NOTE.

Since the actual construction of the k-cubes requires thousands of numbers, we decided to send the 22 pages of tables, separately, to anyone who is interested.

Please send your request for the tables to:

Professor Joseph Arkin
Dept. of Math. Sci.
United States Military Academy
West Point, NY 10996-1786

A Few Brief Historical Notes on Perfect Cubes.

(1) In 1888, the first perfect magic cube ever constructed was of order 8, and was placed in *The Memoirs of the National Academy of Science* [1].

(2) Martin Gardner defines a perfect magic cube as follows:

"A perfect magic cube is a cubical array of positive integers from 1 to N^3

such that every straight line of N cells adds up to a constant. These lines include the orthogonals (the lines parallel to an edge), the two main diagonals of every orthogonal cross section and the four space diagonals.

The constant is

$$(1+2+3+\dots+N^3)/N^2 = \frac{1}{2}(N^4+N)" \quad [2].$$

(3) E.G. Straus, in 1976, in a private letter to Arkin, described how he constructed a $7 \times 7 \times 7$ perfect magic cube. This may be the lowest possible order of a perfect Latin 3-cube [3].

(4) In 1985, Arkin superimposed 6 orthogonal Latin cubes of order 7 to form 20 separate Latin 3-cubes [4].

(5) A perfect 4-dimensional hypercube of order 7 was constructed at West Point in 1989 [5].

(6) A perfect, associative and doubly magic Supercube was constructed at West Point in 1990 [6].

Latin k-cube of order n.

A Latin square of order n is an $n \times n$ square in which each of the numbers $0, 1, \dots, n-1$ occurs exactly once in each row and exactly once in each column. For example

01	012	0123
10	120	1230
	201	2301
		3012

are the Latin squares of order 2, 3, 4, respectively. Two Latin squares of order n are orthogonal, when one is superimposed on another, every ordered pair $00, 01, \dots, n-1 \ n-1$ occurs. Thus

012	012	00 11 22
120 and 201	superimpose to	12 20 01
201	120	21 02 10

and therefore are orthogonal squares of order 3. A set of Latin squares of orders n is orthogonal if every two of them are orthogonal. As an example the 4×4 square of triples

000	111	222	333
123	032	301	210
231	320	013	102
312	203	130	021

represents three mutually orthogonal squares of order 4 since each of the 16 pairs 00,01,...,33 occurs in each of the three possible positions among the 16 triples.

We can generalize all these concepts to $n \times n \times n$ cubes and cubes of higher dimensions. A Latin cube of order n is an $n \times n \times n$ cube (n rows, n columns and n files) in which the numbers $0, 1, \dots, n-1$ are entered so that each number occurs exactly once in each row, column and file. If we list the cube in terms of the n squares of order n which form its different levels we can list the cubes

```

01 10      012 120 201
10 01 and 120 201 012
           201 012 120

```

as Latin cubes of order 2 and 3, respectively.

Orthogonality of Latin cubes is the following relation among three Latin cubes: three Latin cubes of order n are orthogonal if, when superimposed, each ordered triple $000, 001, \dots, n-1 \ n-1 \ n-1$ will occur. For example the pair of 3×3 Latin squares

```

00 11 22
12 20 01
21 02 10

```

leads to the four $3 \times 3 \times 3$ cubes

```

a: 012 120 201
   120 201 012
   201 012 120

```

```

b: 012 120 201
   201 012 120
   120 201 012

```

```

c: 021 102 210
   210 021 102
   102 210 021

```

```

d: 021 102 210
   102 210 021
   210 021 102

```

Superimposed these lead to a cube of quadruples in three levels with I over II over III.

abcd:

```
0000 1122 2211 1111 2200 0022 2222 0011 1100
1221 2010 0102 2002 0121 1210 0110 1202 2021
2112 0201 1020 0220 1012 2101 1001 2120 0212
```

I

II

III

where each ordered triple occurs in every one of the four possible positions in the quadruples.

Note. We define a cube of triples (say abc) where each ordered triple occurs in some order in the 27 cells of the three levels (I over II over III) of the cube abc as a *Latin 3-cube of order 3*.

In this paper we have constructed an orthogonal 4-cube of order 8. This construction consists of eight Latin 4-cubes.

To individually construct each one of the eight cubes, we superimpose 4 orthogonal Latin cubes of order 8. Now, each one of the resulting 4096 cells throughout the construction contains four digits, where each ordered quadruple (0000,0001,...,7777) occurs only once in every cell.

A Perfect Construction.

Each of the eight Latin 4-cubes is perfect in the following way (we consider only one of the eight cubes at a time): the sum (31108) of the elements in each minor diagonal is equal to the sum of the elements of a row in each of the 2 directions in each of the respective squares (layers) that make up each Latin 4-cube of order 8. The sum (31108) of the elements of a row in each direction of a cube is equal to the sum of the elements in each of the 4 major diagonals and the sum on all the diagonals of the cube is the same (namely 31108). The sum (31108) of each of the eight major space diagonals throughout the supercube is the same.

The construction of the cube is based on the 3 orthogonal cubes

$$A_{132} = x_1 + 2x_2 - 3x_3,$$

$$A'_{132} = x_1 - 2x_2 - 3x_3,$$

$$A_{132} = x_1 + 3x_2 + 2x_3,$$

where $(x_1, \dots, x_3) = (0, 1, \dots, 7)$ and arithmetic is (mod 8).

Doubly-Magic Properties.

By labeling the four digits in each quadruple ABCD, new entries for the 4-cube can be formed in each and every one of the 64 squares, in columns, by the following six three-digit combinations - ABA, BAB, BDB, DBD, CDC and DCD. With these new entries, each of the 64 squares in the 4-cube not only has in each column the magic sum of 3108 in base 10, but also a doubly magic sum. That is by squaring each new entry (we consider the squares that are made by using the six column entries - ABA, BAB, BDB, DBD, CDC and DCD instead of ABCD, the new entries) in base 10 and summing the eight appropriate numbers, we obtain 1,640,100.

Super Doubly-Magic Properties.

Considering one square at a time (in what follows, we use all four digits in each cell):

- a) the sum of each of the 8 columns and each of the 2 mini-diagonals are the same
- b) squaring each of the four digits in each cell, we find that the sum of each of the 8 columns and each of the 2 mini-diagonals are the same.

Reversed Digit Construction.

In the constructions that follow, each and every row throughout the 4-cube contains reversed digit pairs. For example: label the 4 digits in any cell wxyz, then on the same row there will always be another cell with four digits labeled xwzy.

Note. Each of the 10 sums in (b) above may vary from square to square.

REFERENCES

1. F. A. P. Barnard, "Theory of Magic Squares and of Magic Cubes," in The Memoires of the National Academy of Science, 4(1888):209-270.
2. Martin Gardner, "Mathematical Games," Scientific American (January 1976), pp. 120,122.
3. In a letter dated January, 1976 E. G. Straus sent Joseph Arkin a detailed construction of a $7 \times 7 \times 7$ perfect magic cube written to base 7 with digits 000 to 666.
4. Joseph Arkin, "An extension of E. G. Straus' perfect Latin 3-cube of order 7," Pacific Journal of Mathematics, Vol. 118, No. 2(June 1985), pp. 277-280.
5. Joseph Arkin, David C. Arney, Bruce J. Porter, "A perfect 4-dimensional hypercube of order 7 (the Cameron cube)," Journal of Recreational Mathematics, Vol. 21, No. 2(1989), pp. 81-88.
6. Joseph Arkin, David C. Arney, Lee S. Dewald, Frank R. Giordano, "Supercube", Proceedings of the Fibonacci Assn., Wake Forest Univ., 8/90.

Bibliography

- a) Arkin, J. and Straus, E.G., "Latin k-cubes," Fibonacci Quarterly, Vol. 12, No. 3(October 1974), pp. 288-292.
- b) Arkin, J., V. E. Hoggatt, Jr., and E. G. Strauss, "Systems of Magic Latin k-Cubes," Canadian Journal of Mathematics. 28:6, pp. 1153-1161, 1976.
- c) Arkin, J., Smith, P., "Treblely magic systems in a Latin 3-cube of order eight," Fibonacci Quarterly, Vol. 14, No. 2(April 1976), pp. 167-170.

CONSTRUCTIONS

In the 4 pages that follow we have exhibited the following constructions:

I. In this table, we list the complete reversed digit and magic construction of a Latin 4-cube of order 3. The 81 numbers (0000 through 2222), are distinct from each other.

II. In this table, we display cube 1 of a Latin 4-cube of order 4, which is made up of 4 cubes, cube 1, ..., cube 4. Each of the 4 cubes contain 64 distinct four digit numbers and each of the 4 cubes is made up of reversed digit and magic constructions. The complete Latin 4-cube of order 4 consists of 256 distinct four digit numbers, 0000 through 3333.

III. In this table, we display cube 1 of a Latin 4-cube of order 8, which is made up of 8 cubes, cube 1, ..., cube 8. Each of the 8 cubes contain 512 distinct four digit numbers and each of the 8 cubes is made up of reversed digit, perfect, magic and doubly magic constructions. The complete Latin 4-cube of order 8 consists of 4096 distinct four digit numbers, 0000 through 7777.

IV. In this table, we display cube 1 of a Latin 4-cube of order 10, which is made up of 10 cubes, cube 1, ..., cube 10. Each of the 10 cubes contain 1000 distinct four digit numbers and each of the 10 cubes is made up of reversed digit and magic constructions. The complete Latin 4-cube of order 10 consists of 10000 distinct four digit numbers, 0000 through 9999.

0000 1221 2112
 1122 2010 0201
 2211 0102 1020
 CUBE 1 SQUARE 1

1111 2002 0220
 2200 0121 1012
 0022 1210 2101
 CUBE 1 SQUARE 2

2222 0110 1001
 0011 1202 2120
 1100 2021 0212
 CUBE 1 SQUARE 3

 1112 2000 0221
 2201 0122 1010
 0020 1211 2102
 CUBE 2 SQUARE 1

2220 0111 1002
 0012 1200 2121
 1101 2022 0210
 CUBE 2 SQUARE 2

0001 1222 2110
 1120 2011 0202
 2212 0100 1021
 CUBE 2 SQUARE 3

 2221 0112 1000
 0010 1201 2122
 1102 2020 0211
 CUBE 3 SQUARE 1

0002 1220 2111
 1121 2012 0200
 2210 0101 1022
 CUBE 3 SQUARE 2

1110 2001 0222
 2202 0120 1011
 0021 1212 2100
 CUBE 3 SQUARE 3

0000 2233 3311 1122
2332 0101 1023 3210
3113 1320 0202 2031
1221 3012 2130 0303
CUBE 1 SQUARE 1

2321 0112 1030 3203
0013 2220 3302 1131
1232 3001 2123 0310
3100 1333 0211 2022
CUBE 1 SQUARE 2

3132 1301 0223 2010
1200 3033 2111 0322
0021 2212 3330 1103
2313 0120 1002 3231
CUBE 1 SQUARE 3

1213 3020 2102 0331
3121 1312 0230 2003
2300 0133 1011 3222
0032 2201 3323 1110
CUBE 1 SQUARE 4

0206	4617	1360	5771	3124	7535	2042	6453
6043	2452	7125	3534	5361	1770	4207	0616
5611	1200	4777	0366	6533	2122	7455	3044
3454	7045	2532	6123	0776	4367	1610	5201
4537	0126	5451	1040	7615	3204	6773	2362
2772	6363	3614	7205	1450	5041	0536	4127
1120	5531	0046	4457	2202	6613	3364	7775
7365	3774	6203	2612	4047	0456	5121	1530

CUBE 1 SQUARE 1

1570	5161	0416	4007	2652	6243	3734	7325
7735	3324	6653	2242	4417	0006	5571	1160
4167	0576	5001	1410	7245	3654	6323	2732
2322	6733	3244	7655	1000	5411	0166	4577
5241	1650	4327	0736	6163	2572	7005	3414
3004	7415	2162	6573	0326	4737	1240	5651
0656	4247	1730	5321	3574	7165	2412	6003
6413	2002	7075	3164	5731	1320	4657	0246

CUBE 1 SQUARE 2

3444	7055	2522	6133	0766	4377	1600	5211
5601	1210	4767	0376	6523	2132	7445	3054
6053	2442	7135	3524	5371	1760	4217	0606
0216	4607	1370	5761	3134	7525	2052	6443
7375	3764	6213	2602	4057	0446	5131	1520
2762	6373	3604	7215	1440	5051	0526	4137
4527	0136	5441	1050	7605	3214	6763	2372

CUBE 1 SQUARE 3

2332	6723	3254	7645	1010	5401	0176	4567
4177	0566	5011	1400	7255	3644	6333	2722
7725	3334	6643	2252	4407	0016	5561	1170
1560	5171	0406	4017	2642	6253	3724	7335
6403	2012	7565	3174	5721	1330	4647	0256
0646	4257	1720	5331	3564	7175	2402	6013
3014	7405	2172	6563	0336	4727	1250	5641
5251	1640	4637	0726	6173	2562	7015	3404

CUBE 1 SQUARE 4

6063	2472	7105	3514	5341	1750	4227	0636
0226	4637	1340	5751	3104	7515	2062	6473
3474	7065	2512	6103	0756	4347	1630	5221
5631	1220	4757	0346	6513	2102	7475	3064
2752	6343	3634	7225	1470	5061	0516	4107
4517	0106	5471	1060	4635	3224	6753	2342
7345	3754	6223	2632	4067	0476	5101	1510
1100	5511	0066	4477	2222	6633	3344	7755

CUBE 1 SQUARE 5

7715	3304	6673	2262	4437	0026	5551	1140
1550	5141	0436	4027	2672	6263	3714	7305
2302	6713	3264	7675	1020	5431	0146	4557
4147	0556	5021	1430	7265	3674	6303	2712
3024	7435	2142	6553	0306	4717	1260	5671
5261	1670	4307	0716	6143	2552	7025	3434
6433	2022	7555	3144	5711	1300	4677	0266
0676	4267	1710	5301	3554	7145	2432	6023

CUBE 1 SQUARE 6

5621	1230	4747	0356	6503	2112	7465	3074
3464	7075	2502	6113	0746	4357	1620	5231
0236	4627	1350	5741	3114	7505	2072	6463
6073	2462	7115	3504	5351	1740	4237	0626
1110	5501	0076	4467	2232	6623	3354	7745
7355	3744	6233	2622	4077	0466	5111	1500
4507	0116	5461	1070	7625	3234	6743	2352
2742	6353	3624	7235	1460	5071	0506	4117

CUBE 1 SQUARE 7

4157	0546	5031	1420	7275	3664	6313	2702
2312	6703	3274	7665	1030	5421	0156	4547
1540	5151	0426	4037	2662	6273	3704	7315
7705	3314	6663	2272	4427	0036	5541	1150
0666	4277	1700	5311	3544	7155	2422	6033
6423	2032	7545	3154	5701	1310	4667	0276
5271	1660	4317	0706	6153	2542	7035	3424
3034	7425	2152	6543	0316	4707	1270	5661

CUBE 1 SQUARE 8

9999 3333 4444 5555 6666 7777 8888 9999 1111 2222
 1551 5115 0880 3775 6226 8008 2662 4994 9449 3337
 2112 8448 1221 7557 4884 0330 5775 3003 4949 9669
 3223 9009 5665 2332 8118 6554 7447 1881 4774 0990
 4334 7997 9779 1001 3443 5225 0110 8668 2552 8886
 5446 0550 8998 9889 2772 4664 1331 7227 5005 3113
 0660 4224 7117 5995 9559 3883 6006 2442 8338 1771
 7007 2882 6336 8228 1991 9119 4554 0770 3663 5445
 8778 1661 3553 0440 5335 2992 9229 6116 7887 4004
 5885 6776 2002 4114 7667 1441 3993 9339 0220 8558

CUBE 1 SQUARE 1

1155 5511 0088 3377 6622 8800 2266 4499 9944 7733
 9494 4949 6262 5801 3737 2626 7373 0158 1085 8510
 7943 2086 9734 8490 0268 6512 4809 5621 3157 1375
 5731 1625 4379 7513 2946 3497 8080 9264 0808 6152
 0518 8150 1805 9624 5081 4739 6942 2376 7493 3267
 3087 6492 2156 1265 7803 0378 9514 8730 4629 5941
 6372 0738 8940 4159 1495 5261 3627 7083 2516 9804
 8620 7263 3517 2736 9154 1945 0498 6802 5371 4089
 2806 9374 5491 6082 4519 7153 1735 3947 8250 0628
 4269 3807 7623 0948 8370 9084 5151 1515 6732 2496

CUBE 1 SQUARE 2

2211 8844 1122 7755 4488 0033 5577 3300 6699 9966
 6309 3690 4578 8034 7965 5487 9756 1212 2121 0843
 9696 5127 6969 0303 1572 4848 3030 8484 7215 2751
 8964 2481 3750 9846 5697 7305 0123 6579 1032 4218
 1842 0213 2031 6489 8124 3960 4698 5757 9306 7575
 7125 4308 5217 2571 9036 1752 6819 0963 3480 8694
 4758 1962 0693 3210 2301 8574 7485 9126 5847 6039
 0483 9576 7845 5967 6219 2691 1302 4038 8754 3120
 5037 6759 9304 4128 3840 9216 2961 7895 0573 1482
 3570 7655 1466 1651 0753 6129 5214 2041 4752 5267

CUBE 1 SQUARE 3

3322 9900 5566 2233 8811 6655 7744 1188 4477 0099
 4187 1478 8741 9650 2093 7814 0239 5326 3562 4905
 0479 7564 4097 6185 5746 8901 1658 9810 2323 3232
 9090 3812 1238 0909 7474 2183 6565 4747 5656 8321
 5906 6325 3652 4817 9560 1098 8471 7234 0189 2743
 2563 8181 7324 3742 0659 5236 4907 6095 1818 9470
 8231 5096 6475 1328 3182 9740 2813 0569 7904 4657
 6815 0749 2903 7094 4327 3472 5186 8651 9230 1568
 7654 4237 9180 8561 1908 0329 3092 2473 6745 5816
 1748 2653 0819 5476 6235 4567 9320 3902 8091 7184

CUBE 1 SQUARE 4

4433 7799 9977 1100 3344 5522 0011 8866 2255 6688
 2865 8256 3014 7529 1680 0341 6108 9437 4973 5792
 6258 0971 2685 5862 9017 3794 8526 7349 1430 4103
 7689 4343 8106 6798 0251 1860 5972 2015 9527 3434
 9797 5432 4523 2345 7979 8686 3254 0101 6868 1010
 1970 3864 0431 4013 6528 9107 2795 5682 8346 7259
 3104 9687 5252 8436 4863 7019 1340 6978 0791 2525
 5342 6018 1790 0681 2435 4253 9867 3524 7109 8976
 0521 2105 7869 3974 8796 6438 4683 1250 5012 9347
 8016 1520 6348 9257 5102 2975 7439 4793 3684 0861

CUBE 1 SQUARE 5

6544 0055 5899 9598 2277 4466 1133 7722 5500 3311
 5720 7502 2137 0465 9318 1273 3581 6649 6894 4056
 3501 1892 5310 4724 8139 2057 7462 0275 9648 6984
 0315 6274 7982 3051 1503 9728 4896 5130 8469 2647
 8059 4646 6464 5270 0895 7312 2507 1983 3721 9138
 9898 2727 1643 6134 3461 8989 5050 4316 7272 0505
 2987 8319 4506 7642 6724 0135 9278 3891 1053 5466
 4276 3131 9058 1313 5640 6504 8729 2467 0985 7892
 1463 5980 0725 2897 7052 3641 6314 9508 4136 8275
 7132 9468 3271 8509 4986 5890 0645 6054 2317 1723

CUBE 1 SQUARE 6

0066 4422 7711 5599 9955 3388 6600 2244 8833 1177
 8243 2834 9605 4382 5179 6950 1597 7061 0716 3428
 1837 6710 8173 3248 7601 9425 2384 4952 5069 0596
 4172 0956 2594 1427 6830 5249 3718 8603 7381 9065
 7421 3068 0386 8953 4712 2174 9835 6590 1247 5609
 5719 9245 6060 0606 1387 7591 8423 3178 2954 4832
 9595 7171 3838 2064 0246 4602 5959 1717 6420 8383
 3958 1607 5429 6170 8063 0836 7241 9385 4592 2714
 6380 8593 4242 9715 2424 1067 0176 5839 3608 7951
 2604 5389 1957 7831 3598 8713 4062 0426 9175 6240

CUBE 1 SQUARE 7

7700 2289 6633 8822 1199 9911 4455 0077 3366 5544
 3076 0367 1459 2918 8542 4195 5824 6703 7630 9281
 5364 4635 3546 9071 6453 1289 0917 2198 8702 7820
 2548 7190 0627 5284 4365 8072 9631 3456 6913 1709
 6283 9701 7910 3196 2658 0547 1369 4825 5074 8452
 8632 1079 4705 7450 5914 6823 3286 9541 0197 2368
 1829 6543 9361 0707 7070 2458 8192 5634 4285 3916
 9191 5454 8262 4545 3706 7360 6073 1919 2828 0637
 4915 3826 2078 1639 0287 5704 7540 8362 9451 5193
 0407 8912 0755 6520 3521 2421 2768 3283 1545 4572

CUBE 1 SQUARE 8

8877 1166 3355 0044 5533 2299 9922 6611 7788 4400
 7618 6781 5923 1296 0404 9532 4040 3875 8357 2169
 4780 9352 7406 2619 3925 5163 6291 1536 0874 8047
 1406 8537 6041 4160 9782 0614 2359 7928 3295 5873
 3165 2879 8297 7538 1356 6401 5783 9042 4610 0924
 0354 5613 9872 8527 4290 3045 7168 2109 6531 1786
 5043 3405 2789 6871 8317 1926 0534 4350 9162 7298
 2539 4920 0184 9402 7878 8787 3615 5293 1046 6351
 9292 7048 1616 5353 6161 4870 8407 0784 2929 3535
 6921 0294 4530 3785 2049 7358 1876 8167 5403 9612

CUBE 1 SQUARE 9

5588 6677 2200 4411 7766 1144 3399 9933 0022 8855
 0932 9023 7396 6147 4851 3769 8415 2580 5208 1674
 8025 3209 0852 1934 2390 7676 9143 6767 4581 5418
 6857 5768 9413 8675 3029 4931 1204 0392 2140 7586
 2670 1584 5148 0762 6207 9853 7026 3419 8935 4391
 4201 7936 3589 5398 8145 2410 0672 1854 9763 6027
 7416 2850 1024 9583 5938 6397 4761 8205 3679 0142
 1764 8395 4671 3859 0582 5028 2930 7146 6417 9203
 3149 0412 6937 7206 9673 8585 5858 4021 1394 2760
 9393 4141 8765 2020 1414 0202 6587 5678 7856 3939

CUBE 1 SQUARE 10

ANALYTIC ROOTS OF THE PERIOD THREE QUADRATIC RECURSION POLYNOMIAL

Harry J. Auvermann

U. S. Army Atmospheric Sciences Laboratory
White Sands Missile Range, New Mexico 88002-5501

ABSTRACT. This paper is concerned with stable points of iterates of the function $F(d,z) = d - z^2$. The number of these stable points changes as the real parameter d varies from $-1/4$ to 2 . The number of stable points is termed the period. Period one stable points are roots of the polynomial that result from substituting z for $F(d,z)$ in the above. Two applications of $F(d,z)$ produce a fourth order polynomial. Period two stable points are roots of this polynomial that are easily obtained. Three applications of $F(d,z)$ produce an eighth order polynomial. Period three stable points are the roots of this polynomial. Two of these roots are known from analysis of the lower iterates. Solution of a sixth order polynomial then determines the period three stable points. An analytic solution to the period four recursion polynomial was reported at last year's conference. We apply the method of the former paper to the period three case and show how the application must be changed for a period that is a prime number.

INTRODUCTION. Motivation for this work has been covered in a previous paper (Auvermann, 1992a). Briefly, the interest arises because transition from order to disorder, similar to the transition of a fluid from laminar to turbulent flow, has been observed in mathematical expressions such as one-dimensional maps, an example being the recursion expression

$$(1) \quad z_{k+1} = d - z_k^2.$$

The parameter d in the mathematical process corresponds to the Reynolds number in the fluid flow process. Corresponding to the random-like samples of the local velocity in the flow are the iterates z_k of the mathematical process. If d is less than a certain value, called the accumulation point, stable points occur (Feigenbaum, 1978). Stable points are repeating numbers in the sequence z_k . This sequence is termed a limit cycle, and the number of points in the cycle is termed the period. If d is larger than the accumulation point, there are isolated intervals wherein stable points occur (Berge', 1984). Between these isolated intervals are intervals of chaos similar to fully developed turbulence. That is, the iteration sequence never repeats itself and the values depend upon the starting point. This similarity between iterate sequences and random processes is the reason for the intense interest in the mathematics of one-dimensional maps and limit cycles.

The condition where each point is the same stable point is analyzed by substitution of z_k for z_{k+1} on the left side of equation (1) and solving for the roots of the resulting polynomial. Bifurcation occurs here in the sense that for larger values of d , the stable points repeat every second iteration. This sequence of two is termed a period two limit cycle. The former case is termed a period one limit cycle. In this paper our attention will be confined to period three limit cycles. Prior work on period three stable points and the general applicability of quadratic recursion has been covered in the former paper (Auvermann, 1992a). Abel (1829) has shown that solutions of polynomials of this type can be reduced

to the solution of polynomials whose order is the same as the period. Abel also gave a method by which all such lower order polynomials may be solved. Netto (1898) has shown how the method of Lagrange resolvents can be used to solve for period three and period four roots.

The new results we report are an alternate method for obtaining the roots of a period three quadratic recursion polynomial. This method, having only been applied previously (Auvermann, 1992a) to period four and now to period three, is not as general as the methods of Abel and Netto. However, it is simpler than that of Netto and serves to accomplish the reduction of the full recursion polynomial to the polynomials solved in general by Abel. Some comments on extension of our method to other periods are contained in the conclusions section below. Much of the algebra necessary to show the results presented here has been left out. More details are contained in a companion government report (Auvermann, 1992b).

The following list contains the essential elements of the remaining notation used. The symbol n is used for the period of a particular limit cycle.

$P_n(d,z)$ = period n recursion polynomial of order 2^n

$E_n(d,z)$ = a factor of $P_n(d,z)$ such that $P_n(d,z) = P_1(d,z) E_n(d,z)$

$Z(d,n,m)$ = m th ($m = 1, 2, \dots, 2^n$) stable point of $P_n(d,z)$

D_n = threshold of d where the roots of $P_n(d,z)$ become real

$j = (-1)^{1/2}$.

$E_n(d,z)$ is called here the reduced polynomial. For periods that are not prime numbers, roots of some lower period polynomials are also roots of the higher period polynomial. When all of these lower order polynomials are factored out, the remaining polynomial is called the primitive polynomial. Using period eight as an example, the primitive polynomial is defined as

$H_8(d,z)$ = a factor of $P_8(d,z)$ such that $P_8(d,z) = P_4(d,z) H_8(d,z)$.

Primitive polynomials are discussed briefly in the conclusion section. When the period is a prime number, the primitive and reduced polynomials are identical.

EXPRESSIONS ASSOCIATED WITH PERIODS ONE, TWO, AND THREE. In this section, the expressions for the polynomials, stable points, and the thresholds for periods one, two, and three will be developed. The first step is to write out the corresponding stable point polynomials.

For a beginning value of the variable, z_0 , and a given parameter, a series of iterates z_k is produced by repeated application of equation (1). If d is greater than D_1 , z_k approaches a fixed point $Z(d,1,m)$ as k increases. This stability occurs when z_{k+1} is equal to z_k in equation (1). The values of z that satisfy this condition are the roots of the period one polynomial

$$(2) \quad P_1(d,z) = z^2 + z - d.$$

where the serial number k has been dropped for writing economy. The period two polynomial is obtained by developing the expression for the iterate two later in the sequence. Hence,

$$(3) \quad P_2(d, z) = (z^2 - d)^2 + z - d.$$

From equations (2) and (3), one has

$$(4) \quad P_2(d, z) = P_1(d, z)[P_1(d, z) - 2z + 1],$$

$$(5) \quad P_2(d, z) = P_1(d, z) E_2(d, z).$$

Equations (4) and (5) serve to define reduced polynomial $E_2(d, z)$. The period three polynomial is obtained in a similar manner by developing the expression for the iterate three later in the sequence. Hence,

$$(6) \quad P_3(d, z) = [(z^2 - d)^2 - d]^2 + z - d.$$

From equations (2) and (6)

$$(7) \quad P_3(d, z) = P_1(d, z)\{[P_1(d, z)]^3 - 2(2z - 1)[P_1(d, z)]^2 + (4z^2 - 6z - 1)P_1(d, z) - 2z(2z - 1) + 1\},$$

$$(8) \quad P_3(d, z) = P_1(d, z) E_3(d, z).$$

Equations (7) and (8) serve to define reduced polynomial $E_3(d, z)$ and show that $P_1(d, z)$ is a factor of $P_3(d, z)$.

Roots of polynomial $P_1(d, z)$ of equation (2) are given by

$$(9) \quad Z(d, 1, 1) = 1/2[-1 + (1 + 4d)^{1/2}],$$

$$(10) \quad Z(d, 1, 2) = 1/2[-1 - (1 + 4d)^{1/2}].$$

$Z(d, 1, 1)$ is a stable fixed point and $Z(d, 1, 2)$ is an unstable fixed point (Feigenbaum, 1983). From equation (8) we see that two roots of the period three polynomial $P_3(d, z)$ are the same as the roots of $P_1(d, z)$. The remaining period three roots are obtained by solving $E_3(d, z)$ from equation (7). The period three reduced polynomial fully expanded is

$$(11) \quad E_3(d, z) = z^6 - z^5 + (1 - 3*d)*z^4 - (1 - 2*d)*z^3 + (1 - 3*d + 3*d^2)*z^2 - (1 - 2*d + d^2)*z + (1 - d + 2*d^2 - d^3).$$

The thresholds will now be considered. For equations (9) and (10) to be real, the radical must be nonnegative. This condition on d defines the period one threshold.

$$(12) \quad D_1 = -1/4.$$

Similar conditions define the thresholds for period two and period four (Auvermann, 1992a) as $D_2 = 3/4$ and $D_4 = 5/4$. The accumulation point has been determined numerically (Berge', 1984), which is $D_\infty = 1.4011519$ when transformed to the parameter of recursion relation (1). The period three threshold, not apparent from equation (11), has been determined numerically to be

$$(13) \quad D_3 = 7/4.$$

In the next section we show how this number arises in the algebra of the root determination process. Thus, the period three interval (where the polynomial roots are real and are stable) begins above the accumulation point at $7/4$.

PERIOD THREE ROOT EXPRESSIONS. The objective is to find analytical expressions for the six roots of $E_3(d, z)$. It is instructive to investigate the root behavior when d is zero to determine how the roots are connected to one another. Under the condition $d = 0$, equation (6) becomes

$$(14) \quad P_3(0, z) = z^8 + z = 0.$$

From equation (14) we see that zero is one root. The other roots are determined from $z^7 + 1 = 0$ and are

$$(15) \quad \begin{aligned} Z(0, 3, 1) &= 0 & Z(0, 3, 2) &= -1, \\ Z_\ell(0) &= Z(0, 3, \ell + 2) = e^{j(2\ell - 7)(\pi/7)}, \quad \ell = 1, 2, \dots, 6. \end{aligned}$$

In equation (15) we have used a more convenient index ℓ and will continue so doing in the remainder of this paper. By applying equation (1) to Z_1 in succession, we find the sequence is Z_1, Z_2, Z_4, Z_1 , etc. By applying equation (1) to Z_6 in succession, we find the sequence is Z_6, Z_5, Z_3, Z_6 , etc. For this case ($d = 0$), the two sequences correspond in conjugate pairs. Because the roots are continuous functions of d , we write the relationships for the roots in general as

$$(16) \quad Z_2(d) = d - [Z_1(d)]^2 \quad Z_5(d) = d - [Z_6(d)]^2,$$

$$(17) \quad Z_4(d) = d - [Z_2(d)]^2 \quad Z_3(d) = d - [Z_5(d)]^2,$$

$$(18) \quad Z_1(d) = d - [Z_4(d)]^2 \quad Z_6(d) = d - [Z_3(d)]^2.$$

At this point we apply the method that was successful in solving the period four polynomial (Auvermann, 1992a), but in a different way. In period four for $d = 0$, the roots are in conjugate pairs within the same sequence. Here they are in conjugate pairs across sequences. The method is a change of variables essentially. We write

$$(19) \quad Z_1 = A + B \quad Z_6 = A - B,$$

$$(20) \quad Z_2 = C + D \quad Z_5 = C - D,$$

$$(21) \quad Z_4 = E + F \quad Z_3 = E - F.$$

Here, A, B, C, D, E, and F are the new variables, each a function of d. B, D, and F are imaginary numbers for $0 > d > D_3$, but will change to real above D_3 . They were written as a general number because they will have the same functional form throughout the range of the d parameter. Therefore, the functional dependence has been dropped to provide additional writing economy. Rewriting equations (16), (17), and (18) in the new variables, we have

$$(22) \quad C + D = d - A^2 - 2AB - B^2 \quad C - D = d - A^2 + 2AB - B^2,$$

$$(23) \quad E + F = d - C^2 - 2CD - D^2 \quad E - F = d - C^2 + 2CD - D^2$$

$$(24) \quad A + B = d - E^2 - 2EF - F^2 \quad A - B = d - E^2 + 2EF - F^2$$

Combining appropriately, we find that

$$(25) \quad C = d - A^2 - B^2 \quad D = -2AB,$$

$$(26) \quad E = d - C^2 - D^2 \quad F = -2CD,$$

$$(27) \quad A = d - E^2 - F^2 \quad B = -2EF,$$

and that

$$(28) \quad 1 = -8ACE.$$

The next step is to define the polynomials associated with each root set. This is done as follows.

$$(29) \quad P_+ = (z - Z_1)(z - Z_2)(z - Z_4),$$

$$(30) \quad P_- = (z - Z_6)(z - Z_5)(z - Z_3).$$

Expanding equation (29), we have

$$P_+ = z^3 - (Z_1 + Z_2 + Z_4) z^2 + (Z_1Z_2 + Z_2Z_4 + Z_4Z_1) - Z_1Z_2Z_4$$

$$P_+ = z^3 - [(A + C + E) + (B + D + F)] z^2$$

$$+ [(AC + CE + EA + BD + DF + FB)$$

$$+ (AD + BC + CF + DE + BE + AF)] z$$

$$- [(ACE + ADF + BCF + BDE)$$

$$(31) \quad + (ACF + ADE + BCE + BDF)],$$

$$(32) \quad P_+ = z^3 - (R + S) z^2 + (T + U) z - (V + W).$$

In equation (32), the symbols R, S, T, U, V, and W represent the corresponding combinations of A, B, C, D, E, and F in equation (31). These are

$$(33) \quad R = A + C + E,$$

$$\begin{aligned}
(34) \quad S &= B + D + F, \\
(35) \quad T &= AC + CE + EA + BD + DF + FB, \\
(36) \quad U &= AD + BC + CF + DE + BE + AF, \\
(37) \quad V &= ACE + ADF + BCF + BDE, \\
(38) \quad W &= ACF + ADE + BCE + BDF.
\end{aligned}$$

These particular combinations were chosen because the set S,U,W changes sign in P_- . As a result, we can write

$$(39) \quad P_- = z^3 - (R - S)z^2 + (T - U)z - (V - W).$$

Since the reduced polynomial is the product of P_+ and P_- , we can write

$$\begin{aligned}
(40) \quad E_3(z,d) &= P_+ P_- \\
&= z^6 - 2Rz^5 + (R^2 - S^2 + 2T)z^4 \\
&\quad - 2(RT - SU + V)z^3 + (2RV - 2SW + T^2 - U^2)z^2 \\
&\quad - 2(TV - UW)z + (V^2 - W^2).
\end{aligned}$$

Equating coefficients between equation (40) and equation (11), we find

$$\begin{aligned}
(41) \quad R &= 1/2, \\
(42) \quad (R^2 - S^2 + 2T) &= (1 - 3d), \\
(43) \quad - 2(RT - SU + V) &= -(1 - 2d), \\
(44) \quad (2RV - 2SW + T^2 - U^2) &= (1 - 3d + 3d^2), \\
(45) \quad - 2(TV - UW) &= -(1 - 2d + d^2), \\
(46) \quad (V^2 - W^2) &= (1 - d + 2d^2 - d^3).
\end{aligned}$$

Another relation is obtained from equation (36) above for U by adding and subtracting the appropriate terms as follows

$$\begin{aligned}
&U = AD + AF + AB - AB + CB + CF + CD - CD \\
&\quad + ED + EB + EF - EF, \\
(47) \quad U &= (A + C + E)(B + D + F) - (AB + CD + EF), \\
(48) \quad U &= RS + (1/2)S = S.
\end{aligned}$$

The simplification from equation (47) to equation (48) came about from applying the right members of equations (25), (26), and (27) and equations (33), (34), and (41).

Substituting equations (41) and (48) into equations (42) through (45) (equation (46) is not needed), we obtain

$$(49) \quad U^2 = 2*T + 3*(d - 1/4),$$

$$(50) \quad T - 2*U^2 + 2*V = (1 - 2*d),$$

$$(51) \quad V - 2*UW + T^2 - U^2 = (1 - 3*d + 3*d^2),$$

$$(52) \quad 2*UW = 2*TV - (1 - 2*d + d^2).$$

Substituting equation (49) into equation (50), equation (52) into equation (51), and then the result into the new (50), we obtain

$$(53) \quad T^2 + 2*T*d + d^2 = 1/4.$$

Since equation (53) is a quadratic involving only T, it may be solved using the quadratic formula and then the remaining coefficients of equation (32) are available directly. A bit of caution is required here, however, to insure that the remaining identifications are consistent. Solving equation (49) for T and substituting into equation (53) and using equation (48), we obtain

$$(54) \quad S^4 - 2*(d - 3/4)*S^2 = 7/16 + (3/2)*d - d^2.$$

Completing the square, we obtain

$$(55) \quad [S^2 - (d - 3/4)]^2 = 1.$$

The roots of equation (55) are

$$\begin{aligned} S &= \pm[\pm 1 + (d - 3/4)]^{1/2} \\ &= \pm(d + 1/4)^{1/2} \text{ or} \\ (56) \quad &= \pm(d - 7/4)^{1/2}. \end{aligned}$$

We will choose the positive root from equation (56) because of the known threshold condition (13), but it is interesting that the period one threshold condition has shown up. The implications of this have not been investigated.

$$(57) \quad S = +(d - 7/4)^{1/2} = U.$$

This result may now be used to obtain all the other coefficients by substitution in sign sensitive expressions.

$$(58) \quad T = (1/2)[S^2 - 3*(d - 1/4)] = -(d + 1/2),$$

$$(59) \quad V = (3/2)*T + (2*d - 1/4) = (1/2)*d - 1,$$

$$(60) \quad W = [TV - (1/2)*(1 - d)^2]/S = -d*(d - 7/4)^{1/2}.$$

S, U, and W are proportional to the factor $(d - 7/4)$, which is zero at $d = D_3$, the threshold condition. This threshold condition manifests itself in the recursion polynomial in a remarkably simple way. Equation (32) may now be written with its full d dependence as

$$\begin{aligned} P_+ = z^3 - [(1/2) + (d - 7/4)^{1/2}] * z^2 \\ + [-(d + 1/2) + (d - 7/4)^{1/2}] * z \\ (61) \quad - [(1/2)(d - 2) - d * (d - 7/4)^{1/2}]. \end{aligned}$$

The roots of equation (61) will be derived below. Since the coefficients of equation (61) can be complex in the general case, there is not in general a real root. At threshold the polynomial is

$$(62) \quad P_+(D_3, z) = z^3 - (1/2) * z^2 - (9/4) * z + (1/8).$$

The numerical values of these roots are

$$(63) \quad Z_1(D_3) = -1.30193773,$$

$$(64) \quad Z_2(D_3) = 0.05495813,$$

$$(65) \quad Z_4(D_3) = 1.74697960.$$

Repeated application of equation (1) cycles through equations (63), (64), and (65) as expected. Since the root expressions for d general are so complicated, only one will be written out. The cubic pattern (Abramowitz, 1970) is

$$(66) \quad z^3 + a_2 * z^2 + a_1 * z + a_0 = 0.$$

In terms of these coefficients, four other auxiliary quantities are defined

$$(67) \quad q = (1/3) * a_1 - (1/9) * a_2^2,$$

$$(68) \quad r = (1/6) * (a_1 * a_2 - 3 * a_0) - (1/27) * a_2^3,$$

$$(69) \quad s_1 = [r + (q^3 + r^2)^{1/2}]^{1/3},$$

$$(70) \quad s_2 = [r - (q^3 + r^2)^{1/2}]^{1/3}.$$

The least complicated root is

$$(71) \quad Z = s_1 + s_2 - a_2/3.$$

In the case of equation (61), the pattern coefficients become

$$(72) \quad a_2 = - [(1/2) + (d - 7/4)^{1/2}],$$

$$(73) \quad a_1 = + [-(d + 1/2) + (d - 7/4)^{1/2}],$$

$$(74) \quad a_0 = - [(1/2)(d - 2) - d * (d - 7/4)^{1/2}].$$

The least complicated root, identified with $Z_4(d)$, is then

$$\begin{aligned}
 (75) \quad Z_4(d) = & \left\{ \frac{1}{6} \right\} [1 + (4d - 7)^{1/2}] \\
 & + \left\{ \frac{1}{3 \cdot 2^{1/3}} \right\} \{ [-14 + 12d - (1 + 8d)(4d - 7)^{1/2}] \\
 & \quad - 3^{3/2} [(4d - 7) - 16d^2 + 8d(4d - 7)^{1/2}]^{1/2} \}^{1/3} \\
 & + \left\{ \frac{1}{3 \cdot 2^{1/3}} \right\} \{ [-14 + 12d - (1 + 8d)(4d - 7)^{1/2}] \\
 & \quad + 3^{3/2} [(4d - 7) - 16d^2 + 8d(4d - 7)^{1/2}]^{1/2} \}^{1/3}
 \end{aligned}$$

Substitution of the threshold D_3 for d returns equation (65). $Z_4(1.8)$ is 1.74734295. Three applications of equation (1) return this same number, verifying the analytical choices.

CONCLUDING REMARKS. The author is indebted to Dr. D. M. Giarrusso, then a member of the Mathematical Sciences Institute at Cornell, now at St. Lawrence University, for the location of the early works of Abel (1829) and Netto (1898). The solution method reported in this paper is independent of Lagrange, but of course gives the same expressions for the roots. Abel proves that solutions of polynomials whose roots are connected by a rational expression and repeat after n applications of this expression (our period n quadratic recursion is a special case) have the following property. Given the order of the recursion polynomial is $v = m \cdot n$, v, n, m integers, then this polynomial may be factored into m polynomials of order n . These polynomials I will call root polynomials. Root polynomials are always solvable for any n because the coefficients in them are related to each other by rational functions that can be determined. Further, if m is a product of other integers, say v_1, v_2, \dots, v_j , then root polynomials may be separated by solution of j other polynomials whose orders are the v 's. These polynomials I will call separation polynomials. The following table is constructed for quadratic recursion to reflect this property.

Table 1. Order Properties of Quadratic Recursion Polynomials

<u>Period</u>	<u>Polynomial Type</u>			
	<u>Recursion</u>	<u>Primitive</u>	<u>Root</u>	<u>Separation</u>
1	2	2	2	N/A
2	4	2	2	N/A
3	8	6	3	2
4	16	12	4	3
5	32	30	5	2, 3
6	64	54	6	3, 3
7	128	126	7	2, 3, 3
8	256	240	8	2, 3, 5

As we have seen above for period 3, the separation polynomial (equation (53)) is order 2 in accordance with table 1. For period 4 (Auvermann, 1992a, equation (35)) the separation polynomial was easily found by the method used there and was

indeed order 3. The separation was so straightforward for period 4 that the root polynomials did not have to be solved. Once the separation polynomial was solved, finding the stable points involved only a series of square root extractions. This simple procedure was not used above for period 3, but rather the root polynomial itself was solved. The fact that equation (28) and equation (80) were not used in the solution perhaps indicates that some manipulation of equations (25), (26), (27), and (28) not yet found would result in the identification of equation (53) without resort to the root polynomials.

There is more than one separation polynomial for periods 5 through 8 because m is not prime. The separation polynomials for periods 5, 6, and 7, if they can be written down, are all solvable as indicated in table 1. Period 8 is the lowest with a separation polynomial of order 5, which is not solvable by conventional techniques. Experience with the method of the present paper has not been deep enough to determine if it will aid solution of periods 5, 6, and 7. Certainly, expressions similar to (25), ..., (28) appear for all periods. These facts coupled with the result from Abel that the individual polynomials for each root set can be solved algebraically give substance to the idea that our method can be extended further than period four.

REFERENCES

- Abel, N. H., 1829, "Mémoire sur une Classe Particulière D'Équations Résolubles Algébriquement," Journal für die Reine and Angewendte Mathematik, Crelle, Berlin.
- Abramowitz, M., and I. A. Stegun, ed., 1970, Handbook of Mathematical Functions, National Bureau of Standards AMS 55, U. S. Government Printing Office, Washington, D. C.
- Auvermann, H. J., 1992a, "Analytic Solution of the Period Four Quadratic Recursion Polynomial," Transactions of the Ninth Army Conference on Applied Mathematics and Computing, U. S. Army Research Office, Research Triangle Park, NC 27709-2211.
- Auvermann, H. J., 1992b, "Roots of the Period Three Quadratic Recursion Polynomial," Technical Report (draft), U. S. Army Atmospheric Sciences Laboratory, White Sands Missile Range, NM 88002-5501.
- Berge', P., Yves Pomeau, and C. Vidal, 1984, Order within Chaos, John Wiley and Sons, New York.
- Feigenbaum, M. J., 1978, "Quantitative Universality for a Class of Nonlinear Transformations," J. Stat. Phys., 19, 25(1978).
- Feigenbaum, M. J., 1983, "Universal Behavior in Nonlinear Systems," Nonlinear Dynamics and Turbulence, ed. G.I. Barenblatt, G. Iooss, and D. D. Joseph, Pitman Advanced Publishing Program, Boston, London, and Melbourne.
- Netto, Eugen, 1898, Verlesungen uber Algebren, Teubner, Leipzig.

A GODUNOV SCHEME FOR ELASTO-PLASTICITY

JOHN W. GROVE
BRADLEY J. PLOHR
DAVID H. SHARP
FENG WANG

ABSTRACT. This paper describes a computational scheme for modeling one-dimensional flow of an elasto-plastic material. The method is based upon a conservative formulation for elasto-plasticity in the Eulerian frame; it uses a second-order Godunov scheme. To validate the method, numerical simulations of loading in plate impact problems are compared with experimental results.

1. INTRODUCTION

Metals and other elasto-plastic materials exhibit a variety of interesting and complicated wave patterns when they are subjected to high stresses and strains. For instance, a shock wave can split into a double wave pattern, with an elastic precursor followed by a plastic compression wave. The precise structure of such a wave reflects the constitutive properties of the material, including the elastic response, yield criterion, and rate sensitivity. To properly resolve elasto-plastic phenomena, numerical methods for modeling the flow must be of high quality.

In this article, we describe a numerical method for modeling one-dimensional elasto-plastic flow. The method is based on a formulation of the governing equations as conservation laws in the Eulerian frame; it employs a high-resolution Godunov method. We validate our method by applying it to the problem of loading in high-velocity impact of metal plates. The results are compared with experimental measurements and with other numerical computations. These comparisons show that the method is successful at resolving the flow, and suggests that extensions of our approach to loading-unloading and multidimensional flows will be useful.

Our computational method is based on a fully conservative Eulerian formulation of the equations of motion for an elasto-plastic medium proposed by Plohr and Sharp [13]. By contrast, the numerical methods commonly used for such computations are nonconservative and Lagrangian. Our computational method has several advantages over the more traditional methods. The Eulerian formulation avoids the problems of poor resolution and numerical diffusion caused by the spatially nonuniform grids and frequent remeshing of Lagrangian

1980 *Mathematics Subject Classification* (1985 Revision). 35L65, 65C20, 73E50, 73F30, 73D05.

Key words and phrases. elasto-plasticity, conservation laws, Eulerian framework, Godunov schemes, loading, plate impact.

This work was supported in part by: the U. S. Army Research Office under Grant DAAL03-92-G-0185; the U. S. Army Research Office under Grant DAAL03-91-C-0027 to the Mathematical Sciences Institute of Cornell University, through subcontract to Stony Brook; and the National Science Foundation under Grant DMS-9057429.

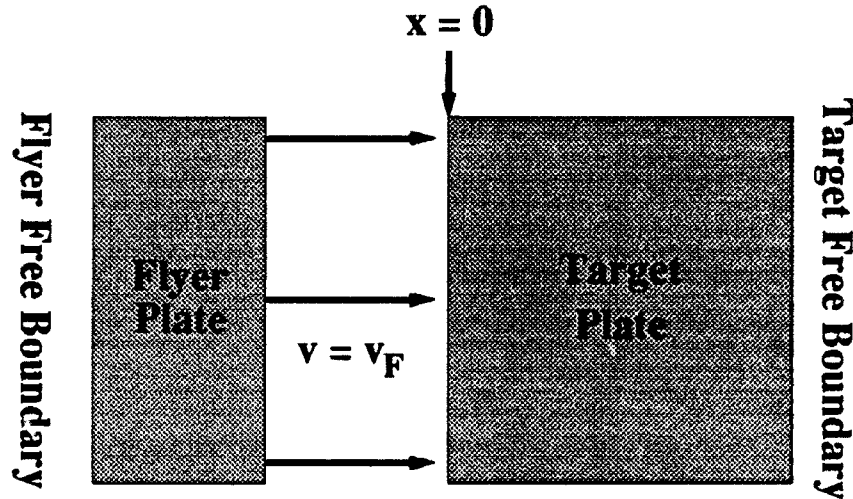


Figure 1: A schematic representation of the impact problem. A flyer plate moving at high velocity collides with a stationary target plate.

methods; instead, we are able to use a fixed spatial grid. The conservative formulation of the flow equations insures that the speeds of the discontinuous waves produced by the interaction are correctly computed. Also, the conservation form of the equation makes it possible to use a high-resolution Godunov method for the computation of the flow. Such methods have been extremely successful in computations of gas dynamical flows, and they are beginning to prove themselves as equally successful in the computation of elastic and plastic waves [18, 2, 5]. In future extensions of this work, we intend to implement front tracking; we expect that tracking will improve the resolution of these waves considerably.

To test our implementation, we have used it to study the high-velocity impact of two tantalum metal plates. A schematic diagram of this experiment is shown in figure 1. A flyer plate moving to the right collides with a stationary target plate; the velocity of the flyer is sufficiently high to produce plastic deformation in both plates. The outer edges of the two plates, the so-called free boundaries, are not constrained. The results of our simulations are compared with the numerical computations of Steinberg and Lund [16] and with the experimental data cited in this reference.

Although the computations presented in the article are only one-dimensional, the conservative Eulerian formulation applies to fully three-dimensional flows. Our implementation has been designed to be extended to treat multidimensional problems. This extension is now being developed, and will be applied to study important problems in elasto-plastic flow, such as the production of shear bands and the multidimensional stability of elastic and plastic waves.

2. ELASTO-PLASTIC DYNAMICS

This section briefly describes the system of nonlinear hyperbolic equations governing the finite deformation of an elastic-plastic material in the Eulerian frame. We use here the fully conservative formulation of these equations recently proposed by Plohr and Sharp [13].

The conservation laws are supplemented by constitutive relations for the elasto-plastic material. In this work, we take the material to be hyperelastic [8, 14], with strain energy of the form appropriate for deformations with small anisotropy [20, 6]; thus the internal energy function is the sum of a hydrostatic term and a small shear term. The plastic flow rule we adopt is that of Lévy-St. Venant (or Prandtl-Reuss): the plastic strain rate is parallel to the deviatoric stress. We impose this rule in the intermediate configuration, so that plastic deformation preserves volume.

The main restriction in the current numerical calculations is that the physical variables depend upon only one spatial variable. In our application to normal impact problems, we also make the simplifying assumption of uniaxial deformations (cf. references [17] and [7]).

Let us now summarize the conservation laws and the constitutive relations. For more on the derivation of the equations, see reference [13].

2.1 Conservation Laws. The deformation of a continuous medium is represented, in the Eulerian picture, by the inverse deformation map ψ^α carrying the medium from its current configuration to its undeformed configuration: $X^\alpha = \psi^\alpha(x^i, t)$, where $i, \alpha = 1, 2, 3$. The evolution of this map is governed by partial differential equations of second order, but these equations can be reduced to first order by introducing the derivatives of ψ^α . The gradient $g^\alpha_i := \partial\psi^\alpha/\partial x^i$ is called the inverse deformation gradient, whereas the time derivative is related to the particle velocity v^i through $\partial\psi^\alpha/\partial t = -g^\alpha_k v^k$. We use g^α_i and v^i as flow variables in place of ψ^α .

The governing conservation laws involve several other quantities that characterize the state of the material and its dynamic response: the Cauchy stress tensor σ^{ij} , the specific internal energy ε , the plastic strain tensor $E_{\alpha\beta}^p$ (measured relative to the undeformed configuration), the plastic source term $\Lambda_{\alpha\beta}$, the strain-hardening parameter κ , and the hardening source term h . These quantities are discussed further below. It is also convenient to introduce the deformation gradient tensor $F^i_\alpha := (g^{-1})^i_\alpha$, the Jacobian $J := \det F$, the mass density $\rho := \rho_0 J^{-1}$ (with ρ_0 being the mass density of the undeformed material), and the specific volume $\tau := 1/\rho$.

Throughout this paper, we assume that the flow variables depend upon only one space variable, say $x := x^1$. In this instance, the principle of continuity (i.e., equality of mixed partial derivatives of ψ^α), the conservation laws of momentum and energy, and the flow rules for the internal plastic variables can be expressed as follows:

$$\frac{\partial}{\partial t} g^\alpha_1 + \frac{\partial}{\partial x} (g^\alpha_k v^k) = 0, \quad (2.1)$$

$$\frac{\partial}{\partial t} g^\alpha_i = 0, \quad i = 2, 3, \quad (2.2)$$

$$\frac{\partial}{\partial t} (\rho v^i) + \frac{\partial}{\partial x} (\rho v^i v^1 - \sigma^{i1}) = 0, \quad (2.3)$$

$$\frac{\partial}{\partial t} [\rho (\frac{1}{2} v_i v^i + \varepsilon)] + \frac{\partial}{\partial x} [\rho (\frac{1}{2} v_i v^i + \varepsilon) v^1 - v_i \sigma^{i1}] = 0, \quad (2.4)$$

$$\frac{\partial}{\partial t}(\rho\kappa) + \frac{\partial}{\partial x}(\rho\kappa v^1) = \rho h, \quad (2.5)$$

$$\frac{\partial}{\partial t}(\rho E_{\alpha\beta}^p) + \frac{\partial}{\partial x}(\rho E_{\alpha\beta}^p v^1) = \rho \Lambda_{\alpha\beta}. \quad (2.6)$$

The principle of continuity implies that the components g^α , for $\alpha = 1, 2, 3$ and $i = 2, 3$ are independent of time. For simplicity, we assume that the initial state is unstrained, i.e., $g|_{t=0}$ is the identity matrix. Therefore, for all time,

$$g^1_2 = g^1_3 = g^2_3 = g^3_2 = 0, \quad g^2_2 = g^3_3 = 1. \quad (2.7)$$

In particular, $g^1_1 = J^{-1}$.

Let us denote by U the 14-component vector of conserved quantities

$$U := (g^\alpha_1, \rho v^i, \rho(\frac{1}{2}v_i v^i + \varepsilon), \rho\kappa, \rho E_{(\alpha)}^p)^T, \quad (2.8)$$

with $\alpha, i = 1, 2, 3$. Here $E_{(\alpha)}^p$ represents Voigt indexing of the six independent components of $E_{\alpha\beta}^p$ (the indices $(\alpha) = 1, \dots, 6$ corresponding to $\alpha\beta = 11, 22, 33, 23, 31, 12$.) Then Eqs. (2.1) and (2.3)–(2.6) form a system of conservation laws

$$\frac{\partial}{\partial t}U + \frac{\partial}{\partial x}H(U) = S(U). \quad (2.9)$$

2.2 Constitutive Relations. Since we assume that the response of the material to deformation is hyperelastic, the thermodynamic variables determine the specific internal energy ε of the material through an equation of state [8] of the form

$$\varepsilon = \hat{\mathcal{E}}(E_{\alpha\beta}, E_{\alpha\beta}^p, \kappa, \eta), \quad (2.10)$$

where

$$E_{\alpha\beta} := \frac{1}{2} [F^k_\alpha F_{k\beta} - \delta_{\alpha\beta}] \quad (2.11)$$

is the (Lagrangian) strain tensor and η the specific entropy. The Clausius-Duhem thermodynamic inequality leads to the identification

$$\sigma^{ij} = \rho F^i_\alpha \frac{\partial \hat{\mathcal{E}}}{\partial E_{\alpha\beta}} F^j_\beta. \quad (2.12)$$

The motivation for the specific choice of energy function in Eq. (2.10) derives from the following microscopic picture. At each point of the material, we imagine cutting out a small neighborhood and relaxing the forces on its surface, recovering an intermediate stress-free configuration. Because of plastic working, this configuration differs from the original Lagrangian configuration. The linear map $(F_p)^\alpha_\alpha$ carrying the Lagrangian configuration to the intermediate configuration is therefore regarded as a measure of local, irrecoverable deformation. This suggests decomposing the deformation gradient as a product [10] of elastic and plastic parts:

$$F^i_\alpha = (F_e)^i_\alpha (F_p)^\alpha_\alpha. \quad (2.13)$$

A GODUNOV SCHEME FOR ELASTO-PLASTICITY

The elastic deformation gradient $(F_e)^i_a$ can be regarded as mapping the intermediate configuration to the Eulerian configuration. For more on the Lagrangian, intermediate, and Eulerian configurations, see reference [14].

In these terms, we identify the Lagrangian plastic strain tensor as

$$E_{\alpha\beta}^p = \frac{1}{2} \left[(F_p)_{\alpha\alpha} (F_p)^{\alpha}_{\beta} - \delta_{\alpha\beta} \right]. \quad (2.14)$$

We also define the elastic strain, measured with respect to the intermediate configuration, to be

$$\bar{E}_{ab}^e := \frac{1}{2} \left[(F_e)_{ka} (\bar{\gamma}_e)^k_b - \delta_{ab} \right]. \quad (2.15)$$

Since we are modeling metals at high strain rates, we assume the material to be isotropic in the intermediate configuration. Therefore the internal energy depends on the principal isotropic invariants of \bar{E}_{ab}^e alone [10]. These isotropic invariants, in fact, can be written as functions solely of $E_{\alpha\beta}$ and $E_{\alpha\beta}^p$ [13]; therefore isotropic energy functions have the form of Eq. (2.10). (From this perspective, the plastic and elastic parts of the deformation gradient appearing in the decomposition (2.13) need never to be introduced.)

Of particular importance is an invariant measure of shear strain. In analogy with the definition used in reference [6], we define such a measure $\bar{\epsilon}_e$ by

$$\bar{\epsilon}_e^2 := \|\text{dev } \bar{E}^e\|^2, \quad (2.16)$$

where the norm $\|A\|$ and deviator $\text{dev } A$ of a 3×3 matrix A are defined by $\|A\|^2 = \text{tr } A^2$ and $\text{dev } A := A - \frac{1}{3} \text{tr } A \, I$, with I being the 3×3 identity matrix. Equivalently, $\bar{\epsilon}_e$ satisfies

$$\bar{\epsilon}_e^2 = \frac{1}{4} \|\text{dev } b_e\|^2, \quad (2.17)$$

where b_e is the elastic Finger tensor

$$b_e := F_e F_e^T = F (I + 2E^p)^{-1} F^T. \quad (2.18)$$

Following Wallace [20] and Garaizar [6], we take the specific internal energy ε to be

$$\varepsilon = \mathcal{E}_h(\tau, \eta) + \tau_0 G(\tau, \eta) \bar{\epsilon}_e^2, \quad (2.19)$$

where the first term represents a hydrostatic contribution to the energy and the second accounts for small shear deformations. Models for the hydrostatic energy \mathcal{E}_h and the shear modulus G are given below. Using the small-shear equation of state (2.19), the identification (2.12) leads to the stress-strain relation

$$\sigma = \left[-p + \tau_0 \frac{\partial G}{\partial \tau} \bar{\epsilon}_e^2 \right] I + J^{-1} G b_e \text{dev } b_e. \quad (2.20)$$

For the hydrostatic component of the energy, \mathcal{E}_h , our computations use a stiffened polytropic equation of state [4, 12]:

$$\mathcal{E}_h(\tau, \eta) := \frac{1}{\gamma - 1} \left(\frac{\tau}{\tau_0} \right)^{-(\gamma-1)} e^{(\gamma-1)\eta/R} (p_0 + p_\infty) \tau_0 + p_\infty \tau. \quad (2.21)$$

Here $\tau_0 = 1/\rho_0$ and p_0 are the specific volume and pressure in the undeformed state, while γ , p_∞ , and R are prescribed constants that characterize the material, γ being dimensionless, p_∞ being a pressure, and R having units of entropy.

In elastodynamics, the pressure p , temperature T , adiabatic bulk modulus K , and Grüneisen coefficient Γ are related to first and second derivatives of the internal energy (see, e.g., reference [20]). In this work, we make the same identification with respect to the hydrostatic energy \mathcal{E}_h :

$$\begin{aligned} p &= -\frac{\partial \mathcal{E}_h}{\partial \tau}, & T &= \frac{\partial \mathcal{E}_h}{\partial \eta}, \\ K &= \tau \frac{\partial^2 \mathcal{E}_h}{\partial \tau^2}, & \Gamma &= -\frac{\tau}{T} \frac{\partial^2 \mathcal{E}_h}{\partial \tau \partial \eta}. \end{aligned} \quad (2.22)$$

By straightforward computations, we obtain the following relations in the undeformed state:

$$RT_0 = (p_0 + p_\infty)\tau_0, \quad K = (p_0 + p_\infty)\gamma, \quad \Gamma = \gamma - 1. \quad (2.23)$$

The bulk modulus K and the Grüneisen coefficient Γ are given in the literature; therefore we can use Eq. (2.23) to calculate R , p_∞ , and γ . Table 1a shows the resulting values for tantalum.

The shear modulus G in our calculations is taken from Steinberg *et al.* [15]:

$$G(\tau, \eta) := G_0 \left[1 + G_p \left(\frac{\tau}{\tau_0} \right)^{1/3} p + G_T (T - T_0) \right], \quad (2.24)$$

with G_0 , G_p , and G_T being material constants and T_0 the temperature in the undeformed state. The constant G_0 is the shear modulus of the material in the unstrained state. See table 1a for the values for tantalum given in reference [15].

2.3 Plastic Flow Rule. To complete the specification of the governing equations, we must define the plastic and hardening source terms. These terms are zero unless the shear stress exceeds a certain threshold, the yield strength. Rather than using the Cauchy stress in the yield criterion, however, we use the stress tensor $\bar{S} := JF_e^{-1}\sigma(F_e^{-1})^T$ corresponding to the intermediate configuration. This is consistent with measuring the shear strain $\bar{\epsilon}_e$ in the intermediate configuration, and it guarantees that plastic deformation does not change the volume (see the discussion below).

One easily sees that $\|\text{dev } \bar{S}\| = J\|\text{dev } \sigma b_e^{-1}\|$. Therefore we interpret the von Mises criterion for elastic flow as

$$\|\text{dev}(\sigma b_e^{-1})\| \leq \sqrt{\frac{2}{3}} Y_0. \quad (2.25)$$

For the static yield strength Y_0 we adopt the model of Steinberg *et al.* [15]:

$$Y_0 := Y_S(\kappa) \frac{G(\tau, \eta)}{G_0}, \quad (2.26)$$

with the strain-hardening part of the yield strength being

$$Y_S(\kappa) := \min \{ Y_A (1 + \beta \kappa)^n, Y_{\max} \}. \quad (2.27)$$

Table 1a						
$\rho_0 (\frac{\text{g}}{\text{cm}^3})$	γ (unitless)	$R (\frac{\text{Mbar}(\text{cm})^3}{\text{kk}})$	p_∞ (Mbar)	G_0 (Mbar)	G_p (Mbar) $^{-1}$	G_T (kK) $^{-1}$
16.6	2.67	0.145	0.72	0.69	1.45	-0.13

Table 1b						
Y_A (Mbar)	β (unitless)	n (unitless)	Y_{\max} (Mbar)	Y_P (Mbar)	U_K (eV)	$k (\frac{\text{eV}}{\text{kk}})$
0.00375	10	0.1	0.011	0.01	0.31	0.086171

Table 1c		
model	$C_1 (\mu\text{s}^{-1})$	C_2 (Mbar μs)
1	0.71	0.12
2	0.71	0.012
3	7.1	0.12
4	7.1	0.012

Table 1: Values of material parameters for tantalum, as used in the numerical simulations. The constants R , p_∞ and γ are computed by using Eq. (2.23), with K and Γ taken from reference [15]. The values for ρ_0 and k can be found in standard physics tables, such as reference [11]. The values for Y_A , β , n , Y_{\max} , Y_P , U_K , G_0 , G_p , and G_T are those used in references [15, 16]. Four sets of values are used for C_1 and C_2 , as indicated in table 1c.

Here Y_A , Y_{\max} , β , and n are material constants. See table 1b for the values corresponding to tantalum.

During plastic loading, where inequality (2.25) is violated, the plastic strain rate tensor is assumed to be parallel to the deviatoric stress, as in the flow rules of St. Venant-Kirchhoff and Prandtl-Reuss. It is suggestive to write the plastic strain rate in the intermediate frame [10, 14] as

$$(L^P \bar{E}^P)_{ab} := (F_p^{-1})^\alpha_a \dot{E}_{\alpha\beta}^P (F_p^{-1})^\beta_b = \tilde{\Lambda} \frac{(\text{dev } \bar{S})_{ab}}{\|\text{dev } \bar{S}\|}, \quad (2.28)$$

where $\tilde{\Lambda}$ is a nonnegative scalar function that vanishes when inequality (2.25) holds. Such a flow rule can be expressed equivalently as

$$\Lambda_{\alpha\beta} = \tilde{\Lambda} \frac{F^T b_e^{-1} \text{dev}(\sigma b_e^{-1}) F}{\|\text{dev}(\sigma b_e^{-1})\|}. \quad (2.29)$$

The flow rule (2.29) (or (2.28)) has the feature that plastic deformation causes no volume change [13], as is usually assumed in metal plasticity. Indeed, the time derivative of $\det F_p = [\det(I + 2E^P)]^{1/2}$ is proportional to the trace of the right-hand of Eq. (2.28), which is zero.

To define $\tilde{\Lambda}$, we adopt the rate-dependent model of Steinberg and Lund [16], making the identification of $\sqrt{\frac{2}{3}}\tilde{\Lambda}$ with the quantity $\dot{\epsilon}_p$ of this reference:

$$\sqrt{\frac{2}{3}}\tilde{\Lambda} := \left(\frac{1}{C_1} \exp \left[\frac{2U_K}{kT} \left(1 - \frac{Y_T}{Y_P} \right)^2 \right] + \frac{C_2}{Y_T} \right)^{-1}, \quad (2.30)$$

where

$$Y_T := \frac{\| \text{dev}(\sigma b_e^{-1}) \| - \sqrt{\frac{2}{3}} Y_0}{\sqrt{\frac{2}{3}} G/G_0} \quad (2.31)$$

is the thermally-activated part of the dynamic yield strength, which measures the extent to which the yield criterion has been exceeded. (Strictly speaking, Y_T is set to zero if it is negative, and to Y_P if it exceeds Y_P .) In the definition of $\tilde{\Lambda}$, the quantities C_1 , C_2 , Y_P , and U_k/k are material constants; values for tantalum are shown in tables 1b and 1c.

Finally, the source term for the hardening law is

$$h := \sqrt{\frac{2}{3}} \tilde{\Lambda} ; \quad (2.32)$$

thus κ represents the accumulated plastic strain.

Hoge and Mukherjee [9] have reported extensive experimental data on tantalum, including plots of Y_T vs. $\dot{\epsilon}_p$ at $T = 0.3$ kK and Y_T vs. T at $\dot{\epsilon}_p = 10^{-10} \mu s^{-1}$. Figure 2 compares these data with the predictions of the models corresponding to the different sets of values for C_1 and C_2 shown in table 1c.

As seen in figure 2, model 4 seems to agree best with the experimental data, particularly in the range of $\dot{\epsilon}_p$ from $10^{-9} \mu s^{-1}$ to $10^{-2} \mu s^{-1}$. This observation is consistent with the results of numerical simulations discussed in section 4 below. Notice that $\dot{\epsilon}_p$ has a maximal, or saturation, value. For model 1 the saturation value is $0.0746 \mu s^{-1}$, whereas for model 4 it is $0.746 \mu s^{-1}$, ten times larger. These values are to be contrasted with the nominal strain rates in the impact problems discussed below, which range from about $0.01 \mu s^{-1}$ to $0.1 \mu s^{-1}$.

2.4 Uniaxial Flow. For the normal impact problem we consider, the governing equations can be simplified because the flow is uniaxial (cf. references [17] and [7]). The inverse deformation gradient has the form $g = \text{diag}(J^{-1}, 0, 0)$ and the velocity vector is $v = (v^1, 0, 0)^T$. Furthermore, the plastic strain is diagonal, the 22 and 33 entries are equal by symmetry, and $\det(I + 2E^p) = 1$ (there being no plastic volume change); therefore we can use a single variable ψ [10, 20] to describe the plastic strain tensor:

$$E^p = \text{diag} \left(\frac{1}{2} (e^{-2\psi} - 1), \frac{1}{2} (e^\psi - 1), \frac{1}{2} (e^\psi - 1) \right) . \quad (2.33)$$

In these terms,

$$b_e = \text{diag} (J^2 e^{2\psi}, e^{-\psi}, e^{-\psi}) \quad (2.34)$$

and

$$\bar{\epsilon}_e^2 = \frac{1}{6} (e^{-\psi} - J^2 e^{2\psi})^2 . \quad (2.35)$$

It proves convenient to ascribe a sign to $\bar{\epsilon}_e$, defining

$$\bar{\epsilon}_e := \frac{1}{\sqrt{6}} (e^{-\psi} - J^2 e^{2\psi}) . \quad (2.36)$$

Then

$$\sigma^{11} = -p + \tau_0 \frac{\partial G}{\partial \tau} \bar{\epsilon}_e^2 - 2\sqrt{\frac{2}{3}} J e^{2\psi} G \bar{\epsilon}_e \quad (2.37)$$

and

$$\Lambda_{11} = -\sqrt{\frac{2}{3}} \text{sgn } \bar{\epsilon}_e e^{-2\psi} \tilde{\Lambda} . \quad (2.38)$$

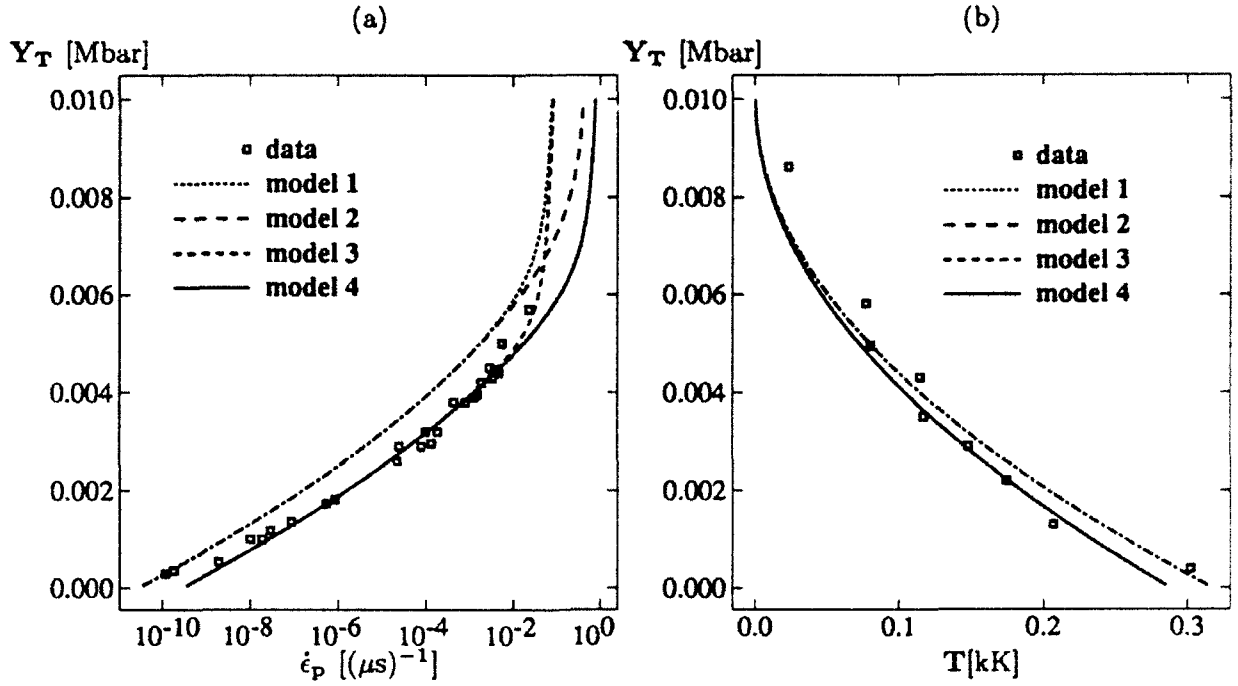


Figure 2: A comparison of constitutive models with experiment. (a) Thermal yield stress vs. plastic strain rate at $T = 0.3$ kK. (b) Thermal yield stress vs. temperature at $\dot{\epsilon}_p = 10^{-10} \mu s^{-1}$; models 1 and 2 are nearly overlapping, as are models 3 and 4.

The conservative system (2.9) can now be simplified to be

$$\frac{\partial}{\partial t} \rho + \frac{\partial}{\partial x} (\rho v^1) = 0, \quad (2.39)$$

$$\frac{\partial}{\partial t} (\rho v^1) + \frac{\partial}{\partial x} (\rho (v^1)^2 - \sigma^{11}) = 0, \quad (2.40)$$

$$\frac{\partial}{\partial t} \left[\rho \left(\frac{1}{2} (v^1)^2 + \varepsilon \right) \right] + \frac{\partial}{\partial x} \left[\rho \left(\frac{1}{2} (v^1)^2 + \varepsilon \right) v^1 - v^1 \sigma^{11} \right] = 0, \quad (2.41)$$

$$\frac{\partial}{\partial t} (\rho \kappa) + \frac{\partial}{\partial x} (\rho \kappa v^1) = \rho h, \quad (2.42)$$

$$\frac{\partial}{\partial t} (\rho \psi) + \frac{\partial}{\partial x} (\rho \psi v^1) = \rho \operatorname{sgn} \bar{\epsilon}_e h, \quad (2.43)$$

together with the constitutive relations (2.19), (2.37), and (2.32).

3. NUMERICAL METHOD

Our computations use a second-order variant of Godunov's method. The basic algorithm for this method is described in references [19, 1, 3, 18]. Since the numerical calculations of this paper are restricted to one space dimension, we describe the method for this case only.

The Godunov type methods require three basic steps: reconstructing the discrete solution from cell averages, computing the interaction of the incoming waves at cell edges using the method of characteristics, and updating cell averages by means of conservative finite differencing.

3.1 Characteristic Analysis. The equations governing smooth elastic-plastic flow can be written in the following characteristic form:

$$\frac{\partial}{\partial t} V + A(V) \frac{\partial}{\partial x} V = \tilde{S}(V), \quad (3.1)$$

where

$$V = (g^{\alpha_1}, v^i, \eta, \kappa, E_{(\alpha)}^p)^T. \quad (3.2)$$

The Jacobian matrix appearing in system (3.1) is

$$A(V) := \begin{pmatrix} v^1 I_{3 \times 3} & g & 0 & 0 & 0 \\ -\rho^{-1} A_g & v^1 I_{3 \times 3} & -\rho^{-1} A_\eta & -\rho^{-1} A_\kappa & -\rho^{-1} A_{E^p} \\ 0 & 0 & v^1 & 0 & 0 \\ 0 & 0 & 0 & v^1 & 0 \\ 0 & 0 & 0 & 0 & v^1 I_{6 \times 6} \end{pmatrix}, \quad (3.3)$$

with

$$\begin{aligned} (A_g)^i_j &:= \frac{\partial \sigma^{i1}}{\partial g^j_1} \Big|_{E^p, \eta, \kappa}, & (A_\eta)^i &:= \frac{\partial \sigma^{i1}}{\partial \eta} \Big|_{E, E^p, \kappa}, \\ (A_\kappa)^i &:= \frac{\partial \sigma^{i1}}{\partial \kappa} \Big|_{E, E^p, \eta}, & (A_{E^p})^{i(\alpha)} &:= \frac{\partial \sigma^{i1}}{\partial E_{(\alpha)}^p} \Big|_{E, \eta, \kappa}. \end{aligned} \quad (3.4)$$

The source term in system (3.1) is

$$\tilde{S}(V) := (0_{1 \times 6}, d/(\rho T), h, \Lambda_{(\alpha)})^T, \quad (3.5)$$

with d denoting the plastic dissipation term

$$d := -\rho \frac{\partial \hat{\mathcal{E}}}{\partial E_{\alpha\beta}^p} \Lambda_{\alpha\beta} - \rho \frac{\partial \hat{\mathcal{E}}}{\partial \kappa} h. \quad (3.6)$$

In order for the system (3.1) to be hyperbolic, $A(V)$ must have real eigenvalues. Evidently, eight of the eigenvalues of $A(V)$ are v^1 , with the corresponding left eigenvectors being appropriate rows of the identity matrix. The remaining eigenvalues are those of the 6×6 matrix

$$B := \begin{pmatrix} v^1 I_{3 \times 3} & g \\ -\rho^{-1} A_g & v^1 I_{3 \times 3} \end{pmatrix}. \quad (3.7)$$

Since

$$\det(B - \lambda I) = -\det\left(-\rho^{-1}A_g g - (v^1 - \lambda)^2 I_{3 \times 3}\right), \quad (3.8)$$

the system (3.1) is hyperbolic if and only if the matrix

$$C := -\frac{1}{\rho}A_g g \quad (3.9)$$

has nonnegative eigenvalues. One can verify that $-\rho C^{ik}$ coincides with the adiabatic acoustic matrix corresponding to the x^1 -direction, i.e., with $a^{ik1} = c^{ik1} + \sigma^{11}\delta^{ik}$, so that C is a symmetric matrix. Therefore there are eight characteristic speeds equal to v^1 , three equal to v^1 plus the positive square roots of the eigenvalues of C , and three equal to v^1 minus these square roots.

Assume that the system (3.1) is hyperbolic. Then we can find a nonsingular matrix L , the rows of which are the left eigenvectors of the matrix C , and a diagonal matrix Π , whose diagonal entries are the square roots of the eigenvalues of C , such that

$$LC = \Pi^2 L. \quad (3.10)$$

We choose L such that $LL^T = I$. The left eigenvectors of B are then determined by the following relation:

$$\begin{pmatrix} -\Pi L f & L \\ \Pi L f & L \end{pmatrix} \begin{pmatrix} v^1 I_{3 \times 3} & g \\ -\rho^{-1} A_g & v^1 I_{3 \times 3} \end{pmatrix} = \begin{pmatrix} v^1 I_{3 \times 3} - \Pi & 0 \\ 0 & v^1 I_{3 \times 3} + \Pi \end{pmatrix} \begin{pmatrix} -\Pi L f & L \\ \Pi L f & L \end{pmatrix}. \quad (3.11)$$

The right eigenvectors can be obtained in a similar manner. Finally, in terms of the 3×8 matrix

$$D := -\frac{1}{\rho}(A_\eta, A_\kappa, A_{Ep}), \quad (3.12)$$

the left and right eigenvectors l_1, \dots, l_{14} and r_1, \dots, r_{14} of the Jacobian matrix $A(V)$ are given by

$$(l_1, \dots, l_{14})^T = \begin{pmatrix} -\Pi L f & L & -\Pi^{-1} L D \\ \Pi L f & L & \Pi^{-1} L D \\ 0 & 0 & I_{8 \times 8} \end{pmatrix} \quad (3.13)$$

and, with $R := L^T$,

$$(r_1, \dots, r_{14}) = \begin{pmatrix} -\frac{1}{2}g R \Pi^{-1} & \frac{1}{2}g R \Pi^{-1} & -g C^{-1} D \\ \frac{1}{2}R & \frac{1}{2}R & 0_{3 \times 8} \\ 0_{8 \times 3} & 0_{8 \times 3} & I_{8 \times 8} \end{pmatrix}. \quad (3.14)$$

These eigenvectors satisfy the normalized biorthogonal condition, i.e., $l_i \cdot r_j = \delta_{ij}$.

3.2 Godunov Scheme. In the numerical scheme, the flow is represented by associating states with points on a grid; each state represents the average of the flow state over the mesh cell centered at the grid point. We choose a fixed Eulerian spatial grid, indexed by i , with spatial increment Δx . To advance the solution from time t_n to t_{n+1} , we use a time step Δt that satisfies the Courant-Friedrichs-Lewy (CFL) stability condition

$$\Delta t = c \frac{\Delta x}{\mu_{\max}}, \quad (3.15)$$

where μ_{\max} is the largest absolute value of the eigenvalues of $A(V)$ and $c < 1$ is a positive constant called the CFL number. In our numerical experiments, $c < 0.5$.

Given the increments Δx and Δt above, the method consists of four steps:

(1) Reconstruction of the flow profile from the cell averages. This step constructs a piecewise linear approximation to $V(x, t_n)$ by determining a slope $\Delta V/\Delta x$, which approximates $\partial V/\partial x$, in each cell. A slope limiter is employed to avoid introducing extraneous local minima and maxima. In our computations, we use the standard van Leer slope limiter [19, 1, 3].

(2) Computation of the half-step left and right states, $V_{i+\frac{1}{2},l}^{n+\frac{1}{2}}$ and $V_{i+\frac{1}{2},r}^{n+\frac{1}{2}}$, at the edges of each cell. These states are found by freezing coefficients in system (3.1) and integrating along characteristics.

(3) Resolution of wave interactions. We solve the Riemann problem for system (2.9) at each cell edge, with left and right states obtained from step (2). We currently use an approximate Riemann solver based on freezing coefficients in system (3.1).

(4) Conservative finite differencing. Using the states $V_{i\pm\frac{1}{2}}^{n+\frac{1}{2}}$ of step 3 to approximate the flux and source terms in system (2.9), the states at time t_{n+1} are obtained from the flux balance relation:

$$U_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x} \left[H \left(\hat{U} \left(V_{i+\frac{1}{2}}^{n+\frac{1}{2}} \right) \right) - H \left(\hat{U} \left(V_{i-\frac{1}{2}}^{n+\frac{1}{2}} \right) \right) \right] \quad (3.16)$$

$$+ \frac{\Delta t}{2} \left[S \left(\hat{U} \left(V_{i-\frac{1}{2}}^{n+\frac{1}{2}} \right) \right) + S \left(\hat{U} \left(V_{i+\frac{1}{2}}^{n+\frac{1}{2}} \right) \right) \right], \quad (3.17)$$

where $\hat{U}(V)$ is the invertible map carrying V to U .

Table 2a	
initial state for flyer	initial state for target
$g = I_{3 \times 3}$	$g = I_{3 \times 3}$
$v^1 = v_F, v^2 = v^3 = 0$	$v^1 = v^2 = v^3 = 0$
$\eta = 0$	$\eta = 0$
$\kappa = 0$	$\kappa = 0$
$E^p = 0_{3 \times 3}$	$E^p = 0_{3 \times 3}$

Table 2b				
test	161	232	310	390
impact velocity (cm/ μ s)	0.0161	0.0232	0.0310	0.0390

Table 2: Initial and parameter values used for the numerical simulations.

4. NUMERICAL RESULTS

We apply our code to the problem of the collision of two tantalum plates. This problem is similar to those presented in references [21, 16, 18]. The initial data and values for the parameters for the constitutive laws are given in tables 2 and 1. The simulation begins as the flyer plate strikes the target plate at $x = 0$. The collision initiates several waves, two moving right and two moving left. The faster wave in each group is called the elastic precursor, and the slower is called the plastic compression wave. When the waves moving right hit the free surface of the target plate, the surface is accelerated. The waves moving left reflect from the free surface of the flyer plate; when they subsequently hit the free surface of the target plate, this surface is decelerated. While the free boundary of the target plate is being accelerated, it is said to be undergoing loading; similarly, the deceleration of this boundary is called unloading. A schematic diagram of the wave structure of the impact problem is shown in figure 3a.

In physical experiments, one typically measures the velocity of the free boundary of the target plate. The velocity profile for the complete loading-unloading process resembles figure 3b. When the elastic-plastic wave group first arrives at the boundary, the elastic precursor raises the velocity to a level that is roughly independent of the impact velocity. As the material yields, the strong plastic wave loads the material to a peak velocity. After a plateau at the peak velocity, which is approximately the same as the impact velocity, waves reflected from the free surface of the flyer plate arrive, causing unloading.

At present, our numerical code is capable only of resolving the loading portion of the collision process. In place of the velocity of the free boundary of the target plate, we calculate the velocity at several fixed points in the target plate. The resulting velocity profiles correspond directly to the free-surface velocity prior to the time of arrival of waves reflected from the free surface of the flyer plate. Thus we can monitor how the shock wave in the target plate splits into a double wave structure.

Figures 4 and 5 show the velocity at three different points plotted vs. time; superimposed in each plot are the results for either different values for C_1 and C_2 , different grid sizes, or

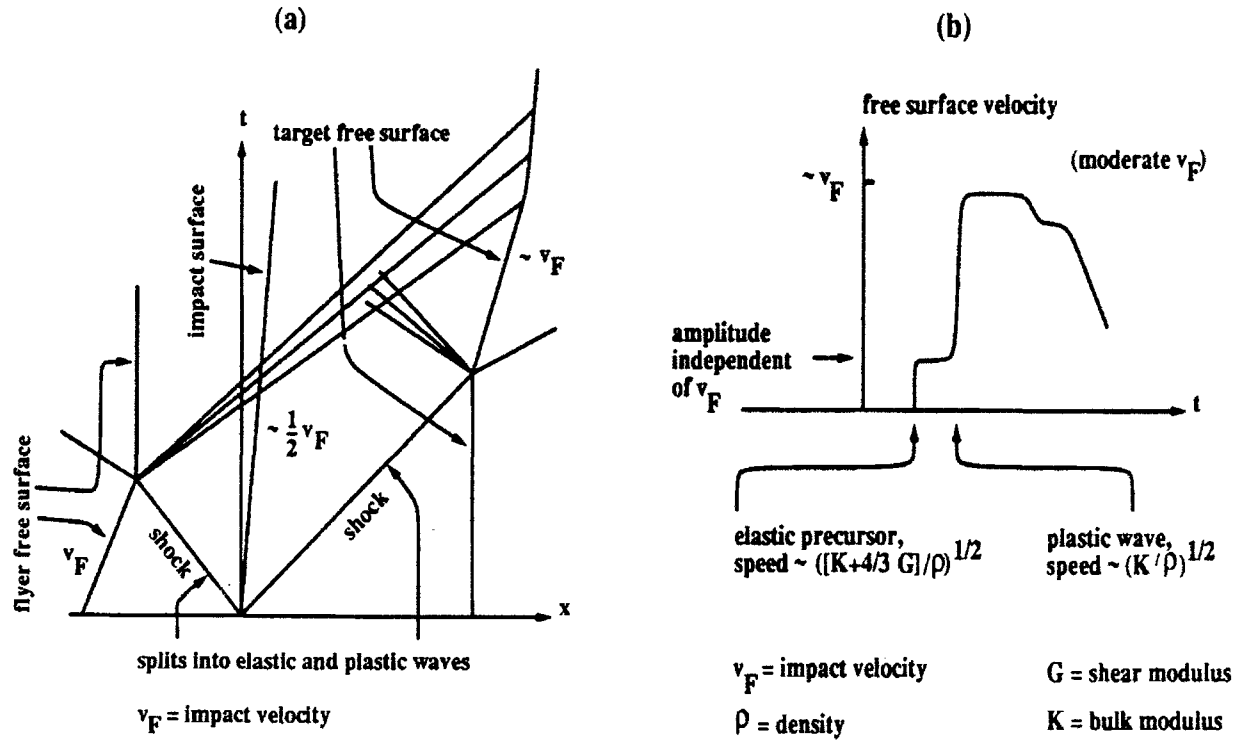


Figure 3: (a) Wave structure of the impact problem. (b) A schematic profile of free surface velocity vs. time.

different impact velocities. In all of these figures except figure 5b, the velocity is measured at the locations $x = 0.1388$ cm, $x = 0.6003$ cm, and $x = 0.96$ cm in the target plate; the corresponding plots appear on the left, in the middle, and on the right.

Since the values of the parameters C_1 and C_2 used in our constitutive model are difficult to determine from fundamental principles, we seek to determine the influence of these parameters on the computed solution. For test 390, figure 4a shows the superposition of the velocity plots for four separate simulations, using the values of C_1 and C_2 indicated in table 1c. Models 1 and 3 produce a similar profile, which shows almost no distinct elastic wave. Models 2 and 4, for which $\dot{\epsilon}_p$ is larger for large Y_T , have more distinct elastic and plastic waves. Model 2 shows an elastic precursor wave with an amplitude about 10% larger than that of model 4. Of the four models used in test 390, model 4 seems to be closest to the experimental results at $x = 0.96$ cm shown in figure 6 of the work of Steinberg and Lund [16]. Figure 4b shows a refinement study for test 390. A grid of 400 cells resolves the problem successfully.

In comparing the results of the four different tests in figure 5a, we observe that the effect of an increased impact velocity is to raise the plateau velocity at the target points. In test 161, the plateau velocity is approximately 1.6×10^{-2} cm/ μ s. Test 232, which uses an impact velocity that is 44% faster, shows a peak velocity at the target points of nearly 2.3×10^{-2} cm/ μ s. Test 310, which uses an impact velocity that is 34% faster than test 232, shows a peak velocity at the target points of almost 3.1×10^{-2} cm/ μ s. Test 390 uses an impact velocity that is 26% higher than test 310; the peak velocity is about 3.9×10^{-2} cm/ μ s. For

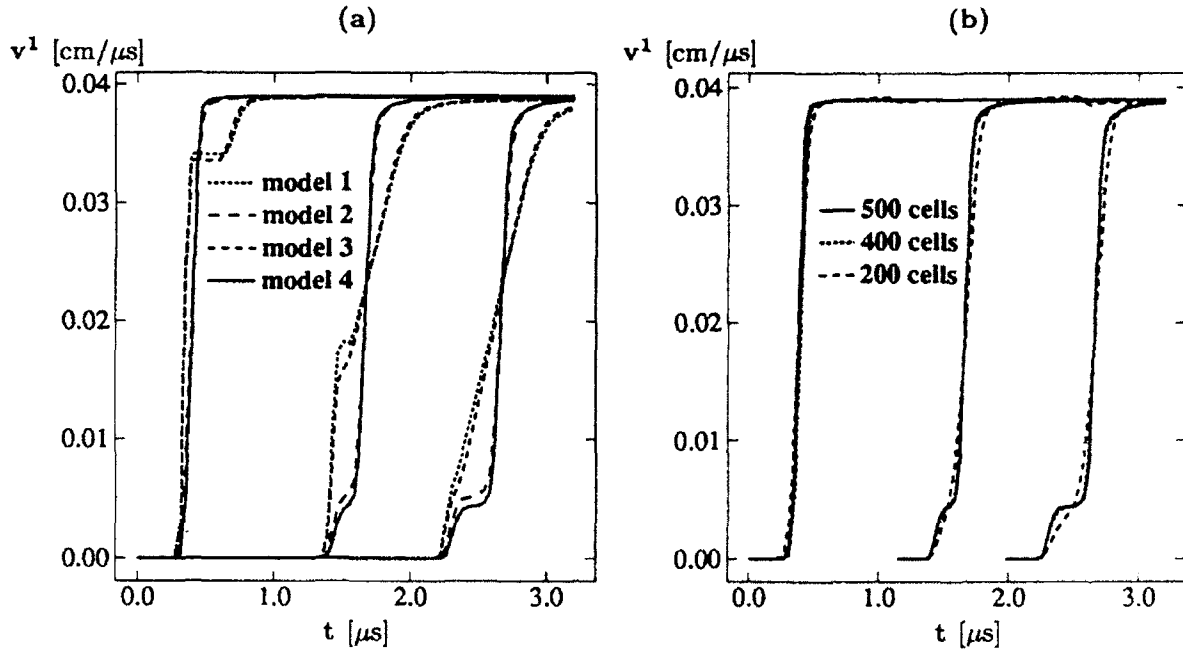


Figure 4: (a) Effect of different values of the parameters C_1 and C_2 on the computed solutions in test 390. (b) Grid refinement study in test 390.

each of the tests, nearly the same peak velocity is obtained at the three points $x = 0.1388$ cm, $x = 0.6003$ cm, and $x = 0.96$ cm. Most importantly, different impact velocities have little influence on the amplitude of the elastic precursor wave.

Finally, figure 5b shows the velocity plots for tests 161, 232, and 390, with parameters C_1 and C_2 coming from model 4. These plots most closely resemble the experimental results shown in figures 4–6 of the work of Steinberg and Lund [16].

5. CONCLUSION

We have described a numerical method for the computation of elasto-plastic flows in one space dimension. The key feature of this numerical method is the use of a conservative Eulerian form of the equations of motion. The equations of motion are solved using a second-order Godunov method. The advantages are the reduction of numerical diffusion in the computed solution and a more uniform distribution of numerical resolution in the computation.

We applied this numerical method to the problem of the high-velocity impact of two metal plates. The results of these computations are in good agreement with computations using other methods and with experimental results. We investigated the effect of varying certain material parameters and observed significant differences in the flow behavior. Our conclusion is that the choice of material parameters corresponding to model 2 of table 1c most closely reproduces the results of experiment.

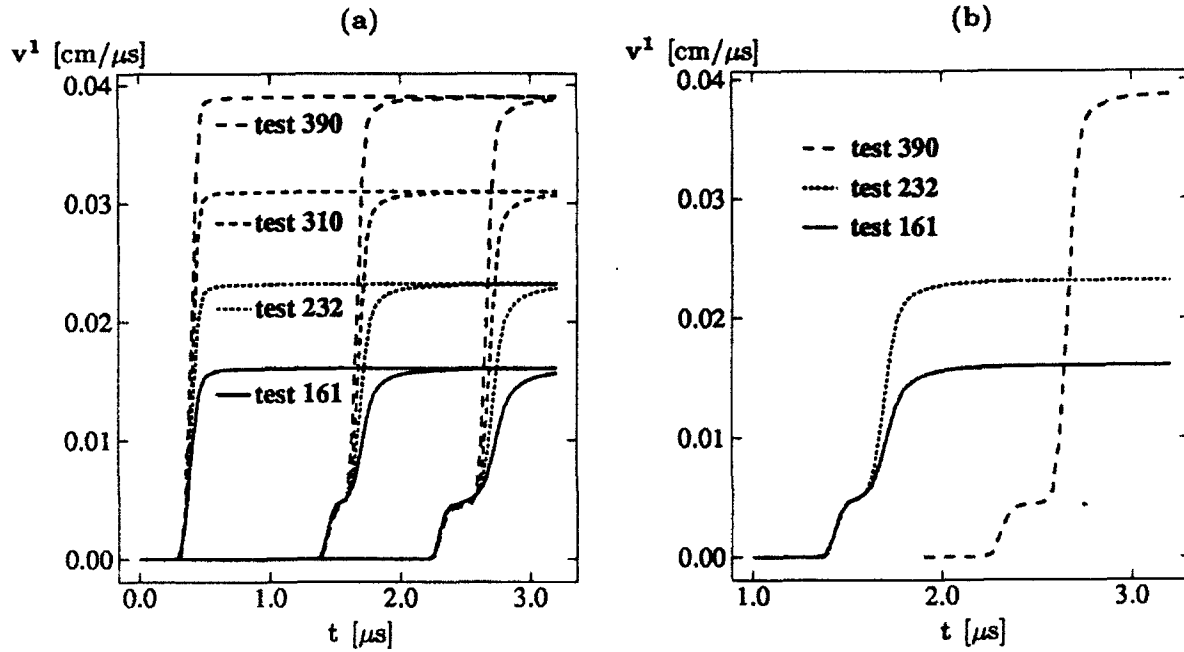


Figure 5: (a) Effect of different impact velocities on the computed solutions. (b) Velocity vs. time for tests 161, 232, and 390.

Future directions of development of our numerical code will include the use of front tracking for the elastic and shear waves in the solid medium, and, more importantly, the extension of the method to flows in two and three space dimensions.

ACKNOWLEDGMENTS

The authors thank James Glimm for his encouragement and insightful comments. We also thank the Army Research Office and the National Science Foundation for their support of this work.

REFERENCES

- [1] J. BELL, P. COLELLA, AND J. TRANGENSTEIN, *Higher order Godunov methods for general systems of hyperbolic conservation laws*, J. Comput. Phys., 82 (1989), pp. 362–397.
- [2] E. BONNETIER, H. JOURDREN, AND P. VEYSSEYRE, *Un modèle hyperélastique-plastique eulérien applicable aux grandes déformations: Quelques résultats 1-d*, tech. rep., Centre d'Etudes de Limeil-Valenton, 1991.
- [3] P. COLELLA, *Multidimensional upwind methods for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 171–200.
- [4] R. COURANT AND K. FRIEDRICHS, *Supersonic Flow and Shock Waves*, Interscience, New York, 1948.
- [5] P. L. FLOCH AND F. OLSSON, *A second-order Godunov method for the conservation laws of nonlinear elastodynamics*, Impact Comput. Sci. Engrg., 2 (1990), pp. 318–354.
- [6] X. GARAIZAR, *The small anisotropy formulation of elastic deformation*, Acta Appl. Math., 14 (1989), pp. 259–268.
- [7] —, *Solution of a Riemann problem for elasticity*, J. Elasticity, 26 (1991), pp. 43–63.
- [8] A. GREEN AND P. NAGHDI, *A general theory of an elastic-plastic continuum*, Arch. Rational Mech. Anal., 18 (1965), pp. 251–281.
- [9] K. HOGE AND A. MUKHERJEE, *The temperature and strain rate dependence of the flow stress of tantalum*, J. Mater. Sci., 12 (1977), pp. 1666–1672.
- [10] E. LEE AND D. LIU, *Finite-strain elastic-plastic theory with application to plane-wave analysis*, J. Appl. Phys., 38 (1967), pp. 19–27.
- [11] C. NORDLING AND J. ÖSTERMAN, *Physics Handbook*, Studentlitteratur, Sweden, 1980.
- [12] B. PLOHR, *Shockless acceleration of thin plates modeled by a tracked random choice method*, AIAA J., 26 (1988), pp. 470–478.
- [13] B. PLOHR AND D. SHARP, *A conservative formulation for plasticity*, Adv. Appl. Math., (1992). To appear.
- [14] J. SIMO AND M. ORTIZ, *A unified approach to finite deformation elastoplastic analysis based on the use of hyperelastic constitutive equations*, Comput. Methods Appl. Mech. Engrg., 49 (1985), pp. 221–245.
- [15] D. STEINBERG, S. COCHRAN, AND M. GUINAN, *A constitutive model for metals applicable at high strain-rate*, J. Appl. Phys., 51 (1980), pp. 1498–1504.

- [16] D. STEINBERG AND C. LUND, *A constitutive model for strain rates from 10^{-4} to 10^6 s^{-1}* , J. Appl. Phys., 65 (1989), pp. 1528–1533.
- [17] Z. TANG AND T. TING, *Wave curves for the Riemann problem of plane waves in simple isotropic elastic solids*, Int. J. Engrg. Sci., 25 (1987), pp. 1343–1381.
- [18] J. TRANGENSTEIN AND P. COLELLA, *A higher-order Godunov method for modeling finite deformation in elastic-plastic solids*, Commun. Pure Appl. Math., 44 (1991), pp. 41–100.
- [19] B. VAN LEER, *Towards the ultimate conservative difference scheme: IV. A new approach to numerical convection*, J. Comput. Phys., 23 (1977), pp. 276–299.
- [20] D. WALLACE, *Thermoelastic-plastic flow in solids*, Tech. Rep. LA-10119, Los Alamos National Laboratory, 1985.
- [21] M. WILKINS, *Calculation of elastic-plastic flow*, Methods Comput. Phys., 3 (1964), pp. 211–263.

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS, STATE UNIVERSITY OF NEW YORK
AT STONY BROOK, STONY BROOK, NY 11794-3600

E-mail: grove@ams.sunysb.edu

DEPARTMENTS OF MATHEMATICS AND OF APPLIED MATHEMATICS AND STATISTICS, STATE UNIVERSITY OF NEW YORK AT STONY BROOK, STONY BROOK, NY 11794-3600

E-mail: plohr@ams.sunysb.edu

THEORETICAL DIVISION, MS-B285, LOS ALAMOS NATIONAL LABORATORY, LOS ALAMOS, NM 87545

E-mail: dhs@t13.lanl.gov

DEPARTMENT OF APPLIED MATHEMATICS AND STATISTICS, STATE UNIVERSITY OF NEW YORK
AT STONY BROOK, STONY BROOK, NY 11794-3600

E-mail: fwang@ams.sunysb.edu

The Korteweg Theory of Cappilarity and the Phase Transition Problems *

Harumi Hattori
Department of Mathematics
West Virginia University
Morgantown, WV 26506

Abstract

In this paper we first summarize the earlier results on the slow motion in the Korteweg theory of cappilarity in the one-dimensional case and show some numerical results. In the multidimensional case we discuss the existence of local solutions to the system of equations for compressible fluids of Korteweg type.

1 Introduction

In order to model the cappilarity effect of materials, Korteweg [12] formulated a constitutive equation for the Cauchy stress that includes density gradients. It turns out that his theory is useful to discuss phase transition problems.

First, we discuss the one-dimensional isothermal motion. In this case the equation we discuss is given by

$$(1.1) \quad u_{tt} = \sigma(u_x)_x + \nu u_{xxt} - \epsilon^2 u_{xxxx}, \quad 0 < x < 1, \quad t > 0.$$

where u is the displacement and u_{xxt} and u_{xxxx} terms represent the viscosity and the cappilarity effects, respectively. Typical boundary conditions come from either a soft loading device or a hard loading device. Although the slow motion occurs in both cases, in this note we discuss the soft loading case only for simplicity. The boundary conditions in this case are given by

$$(1.2) \quad \begin{aligned} u(0, t) &= 0, \quad \sigma(u_x) + \nu u_{xt} - \epsilon^2 u_{xxx}|_{x=1} = P, \\ u_{xx}(0, t) &= 0, \quad u_{xx}(1, t) = 0. \end{aligned}$$

The initial conditions are given by

$$(1.3) \quad u(x, 0) = f(x), \quad u_t(x, 0) = g(x),$$

where $f, g \in H^1(0, 1)$. The boundary conditions (1.2a) show that the stress P is applied at $x = 1$. The boundary conditions (1.2b) are the natural boundary conditions for the corresponding variational problem.

*The author was supported in part by Army Grant DAAL 03-89-G-0088.

In what follows, we assume that σ is given by Fig. 1.1. In this figure $(0, \alpha^*)$ and (β^*, ∞) are called the α -phase and the β -phase, respectively. They correspond to the different phases of the material. The interval (α^*, β^*) is called the spinodal region and physically unstable. We denote by α , δ , and β the values of u_x at the intersections of $y = P$ and $y = \sigma(u_x)$ in the α -phase, the spinodal region, and the β -phase, respectively. The value of P for which areas A and B are equal is called the Maxwell line. We denote by α_M , β_M , and δ_M the values of α , β , and δ , respectively, for which we have the Maxwell line construction.

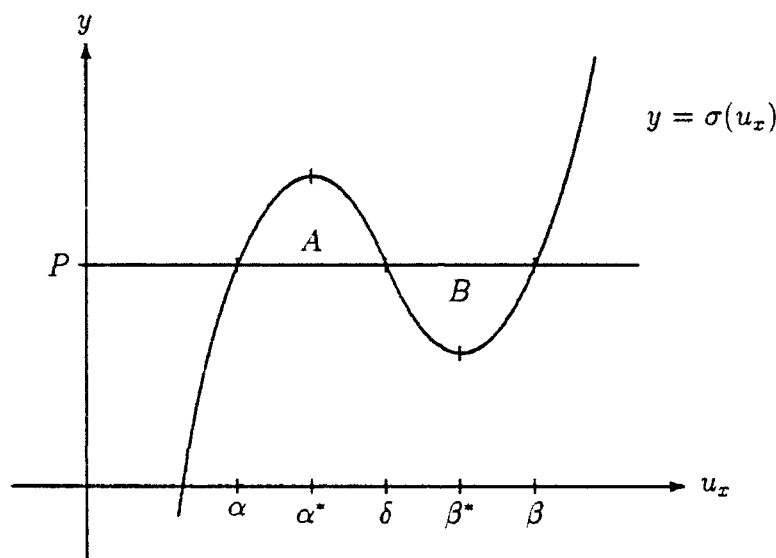


Figure 1.1

The capillarity term was first introduced by Korteweg [12]. Recently, various effects of this term have been discussed. For example, Serrin [15], [16] reconsidered the Korteweg theory and has shown the existence of steady profile connecting the α -phase and the β -phase. Slemrod [17] and Hagan and Slemrod [9] considered the existence of travelling wave solutions. The static problems concerning the soft loading case and the hard loading case have been discussed in [3] and [4], respectively. The dynamical aspects of these loading cases are discussed in Hattori and Mischaikow [10] and Andrews and Ball [1].

In Section 2 we summarize the result in [11] about a slow motion of (1.1) resembling the dynamics of (2.3) discussed in [7], [5], [2], and [8]. In Section 3 we show some numerical examples of slow motions. In Section 4 we discuss the existence of local solutions to the system of equations for two dimensional isothermal motion of compressible fluids of Korteweg type. The higher order terms of density (or the deformation gradient) in the Cauchy stress tensor is not in general compatible with the classical theory of thermody-

namics. Dunn and Serrin [6] introduced the concept of interstitial working and derived the Cauchy stress tensor compatible with the thermodynamics. First, we summarize their results and derive the system of equations. Then, we state the theorem for the existence of local solutions.

2 Slow motions one-dimensional case

In this section we summarize the results in [2], [8], and [11]. Multiply (1.1) by u_t , integrate in x and t , and then integrate by parts using (1.2). After dividing by ϵ , we have,

$$(2.1) \quad E[u](t) + \frac{1}{\epsilon} \int_0^t \int_0^1 \nu u_{xt}^2(x, s) dx ds = E[u](0),$$

where

$$(2.2) \quad E[u](t) = \int_0^1 \left\{ \frac{1}{2\epsilon} u_t^2 + \frac{1}{\epsilon} (W(u_x) - Pu_x) + \frac{\epsilon}{2} u_{xx}^2 \right\} (x, t) dx.$$

In (2.2) $W(u_x)$ is a primitive of σ . For the remainder of the paper we shall assume that $P = \sigma(\alpha_M)$. This implies that $W(u_x) - Pu_x$ will be double-well potentials with equal depth. For the sake of simplicity we shall also assume that $W(u_x) - Pu_x$ is given by $(u_x - 1)^2$. The same conclusions will hold for more general non-linearities.

Observe that (2.1) is similar to that for the parabolic equation

$$(2.3) \quad \nu v_t = \epsilon^2 v_{xx} - (v^3 - v),$$

with either the homogeneous Neumann boundary condition

$$v_x(0, t) = 0, \quad v_x(1, t) = 0$$

or a Dirichlet condition

$$v(0, t) = a, \quad v(1, t) = b, \quad a, b = \pm 1.$$

In particular the energy relation for (2.3) is given by

$$(2.4) \quad E_p[v](t) + \frac{1}{\epsilon} \int_0^t \int_0^1 \nu v_t^2(x, s) dx ds = E_p[v](0),$$

where

$$(2.5) \quad E_p[v](t) = \int_0^1 \left\{ \frac{1}{4\epsilon} (v^2 - 1)^2 + \frac{\epsilon}{2} v_x^2 \right\} (x, t) dx.$$

Now we summarize the results of the slow motions for the parabolic equation. We assume for the initial data of (2.3) that

$$(2.6) \quad w(x) = \lim_{\epsilon \rightarrow 0} v^\epsilon(x, 0)$$

exists as a limit of L^1 norm, where w is a piecewise constant function taking only the values ± 1 , with exactly N discontinuities at $\{x_1, \dots, x_N\}$ and we also assume that the initial data satisfy

$$(2.7) \quad E_p[v^\epsilon](0) \leq Nc_0 + K_2 \exp(-K/\epsilon).$$

Then, we have

Lemma 2.1 *Suppose the initial data for (2.3) satisfy (2.6) and (2.7). Then, for any T satisfying $0 \leq T \leq F\nu\epsilon^s \exp(-K/\epsilon)$, we have*

$$(2.8) \quad \sup_{0 \leq t \leq T} \int_0^1 |v^\epsilon(x, t) - v^\epsilon(x, 0)| dx \leq (FG)^{1/2} \epsilon^{\frac{1}{2}(s+1)}.$$

Next, we summarize the results concerning the slow motions of (1.1). As the form of the energy relation (2.1) resembles (2.4), we can expect to draw the same kind of conclusions for (1.1). For this purpose we rewrite the energy $E_c[u]$ as

$$E_c[u] = E_s[u] + E_p[u_x], \quad E_s[u] = \int_0^1 \frac{1}{2\epsilon} u_t^2(x, t) dx.$$

We assume that the initial data for (1.1) satisfy

$$(2.9) \quad u_x^\epsilon(x, 0) = v^\epsilon(x, 0)$$

and

$$(2.10) \quad E_s[u^\epsilon](0) \leq C \exp(-K/\epsilon).$$

The condition (2.9) is imposed for the sake of simplicity. As long as $u_x(x, 0)$ satisfies (2.6) and (2.7), with v being replaced by u_x , the same conclusion should be obtained.

Lemma 2.2 *Suppose (2.9) and (2.10) are satisfied. Then, for any T satisfying $0 \leq T \leq F_c\nu\epsilon^s \exp(K/\epsilon)$, the solution to (1.1) satisfies*

$$(2.11) \quad \sup_{0 \leq t \leq T} \int_0^1 |u_x^\epsilon(x, t) - u_x^\epsilon(x, 0)| dx \leq (F_c G_c)^{1/2} \epsilon^{\frac{1}{2}(s+1)}.$$

Using Nirenberg's inequality and Lemmas 2.1 and 2.2, we can show

Theorem 2.3 *If $s > 1$, then the difference in the L^∞ norm between u_x^ϵ and v^ϵ is $O(\epsilon^{\frac{1}{6}(s-1)})$ for at least $0 \leq t \leq F_m\nu\epsilon^s \exp(K/\epsilon)$, where $F_m = \min\{F, F_c\}$.*

3 Numerical examples

We give a numerical example of the soft loading case to confirm the results in the previous section. We introduce the transform

$$p = \int_1^x u_t(x, t) dx, \quad q = u_x$$

similar to Pego's [14]. Then, (1.1) becomes

$$(3.1) \quad \begin{aligned} p_t &= \nu p_{xx} - \eta q_{xx} + \sigma(q) - P, \\ q_t &= p_{xx}. \end{aligned}$$

The boundary conditions for p and q become

$$(3.2) \quad \begin{aligned} p_x(0, t) &= 0, & p(0, t) &= 0, \\ q_x(0, t) &= 0, & q_x(1, t) &= 0. \end{aligned}$$

For the initial condition, we consider the case when

$$(3.3) \quad p(x, 0) = 0, \quad q(x, 0) = Cf(x),$$

where C is a parameter representing the magnitude of initial data.

As an example, we consider the case when $\epsilon = 0.01$, $\nu = 1.0$, and the initial data for the parabolic equation and for (3.1) are given, respectively, by

$$\begin{aligned} v(x, 0) &= C(\cos 2\pi x + \cos 9\pi x), \\ p(x, 0) &= 0, \quad q(x, 0) = v(x, 0). \end{aligned}$$

For C we gave the following values:

$$C = 1.0, 0.5, 0.1, 0.01, 0.001.$$

One of the reasons why we change the magnitude of the initial data is to see how this influences the metastable states. We should note that for either choice of C above, the conditions (2.9) and (2.10) are not satisfied. Nevertheless, when $C = 1.0, 0.5, 0.1$, v and q have reached the same metastable state in each case. Here, we show the numerical results of $C = 0.1, 0.01$ only. In Figures 3.1 and 3.2 we show how v and q evolve for $0 \leq t \leq 10$ and then in Figures 3.3 and 3.4 we show the profiles of v and q at $t = 1000$. We use the solid lines for v and the gray lines for q . When these lines overlap, we see only the gray lines. When $C = 0.1$, they agree at least to 10^{-7} at $t = 1000$ and this agreement continues at least until $t = 10000$. In these figures, the values of x should be multiplied by 0.01.

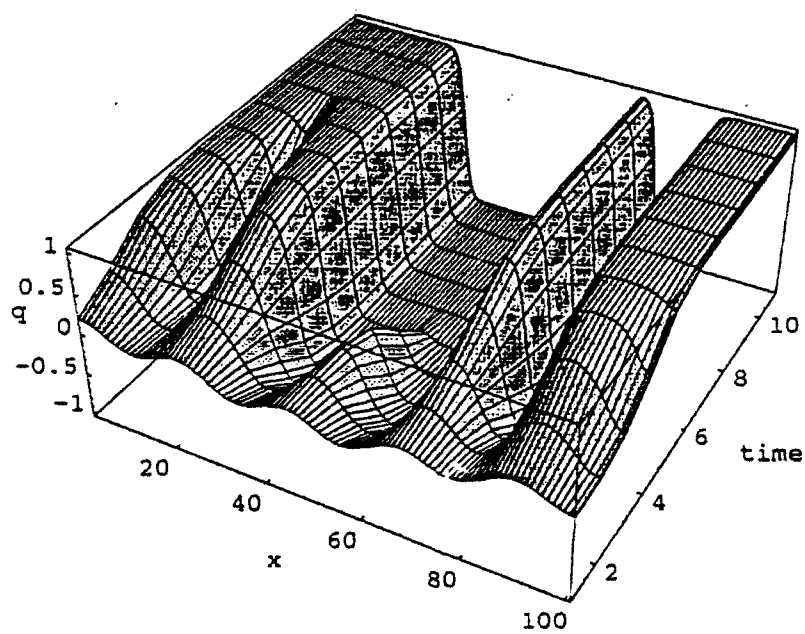
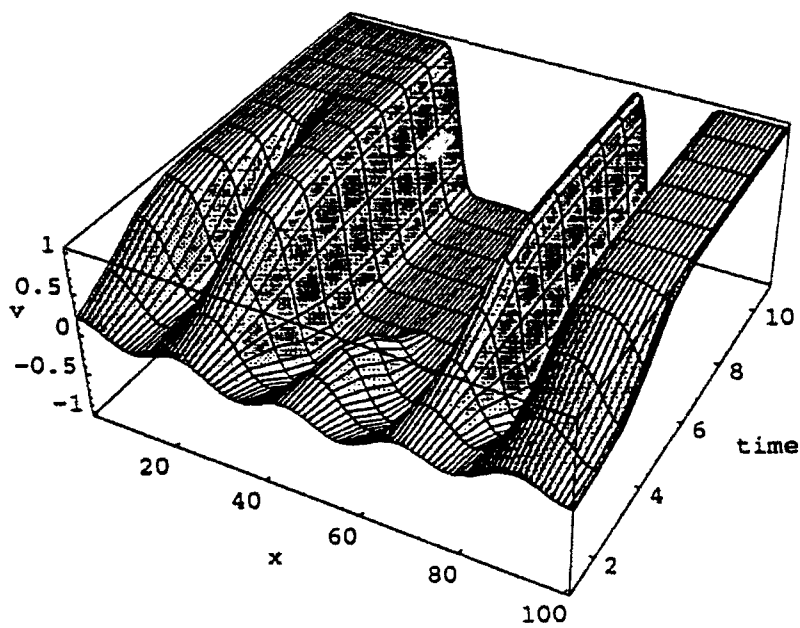


Figure 3.1. v and q for $C = 0.1$, $0 \leq t \leq 10$.

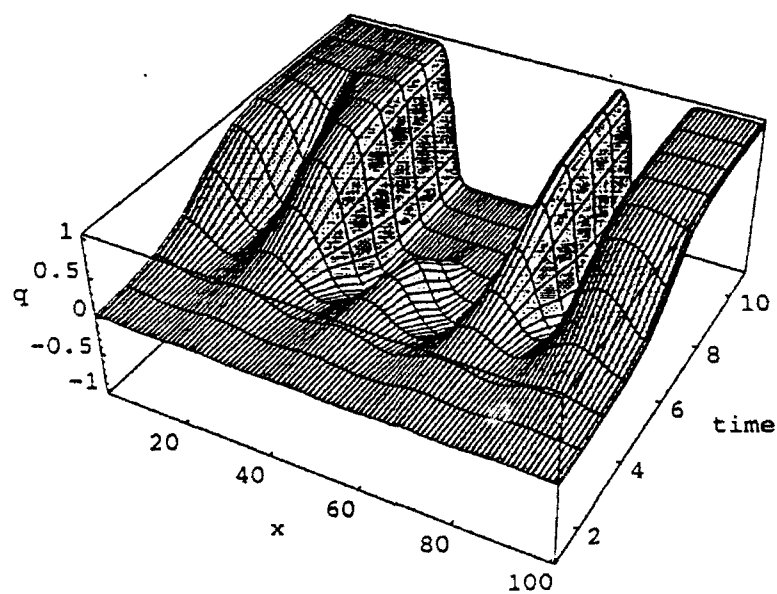
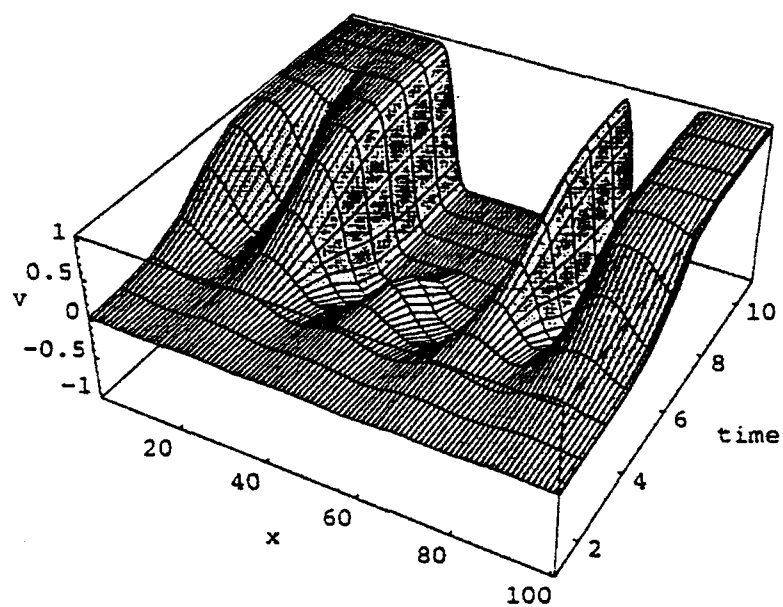


Figure 3.2. v and q for $C = 0.01$, $0 \leq t \leq 10$.

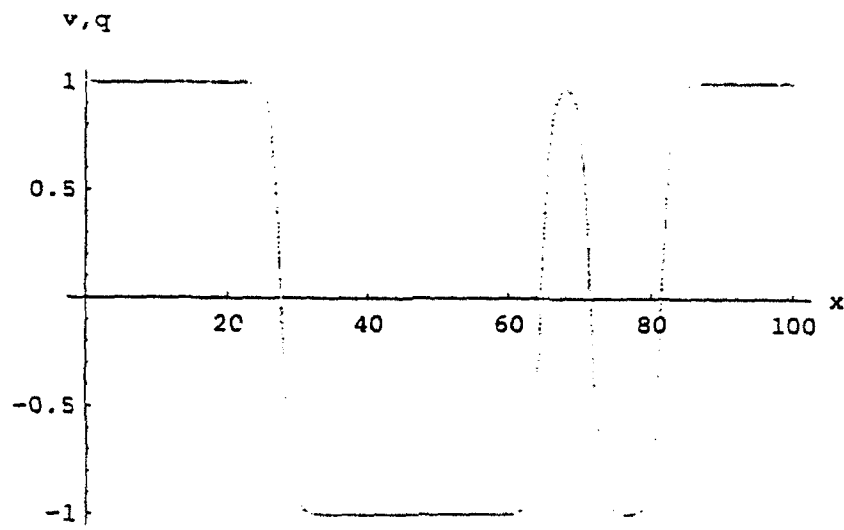


Figure 3.3. v and q at $t = 1000$ for $C = 0.1$.

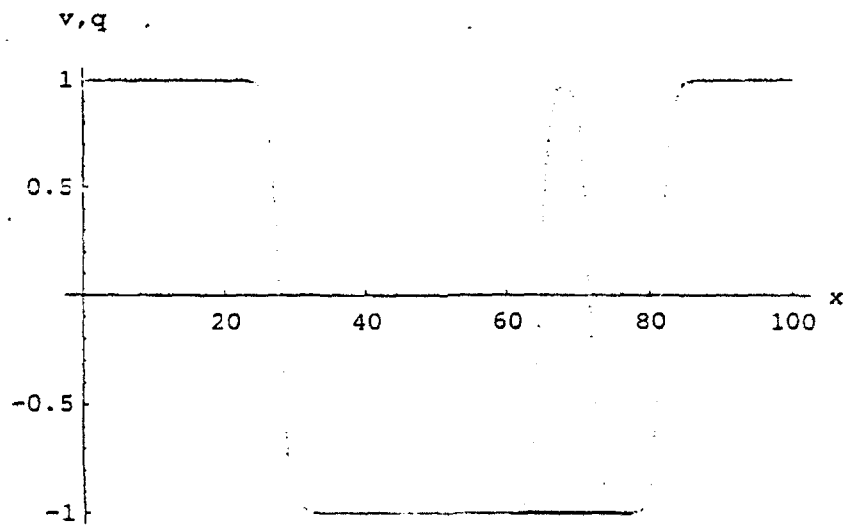


Figure 3.4. v and q at $t = 1000$ for $C = 0.01$.

4 Local existence in multidimensional case

Dunn and Serrin [6] modified the Korteweg theory and derived the following set of equations for the conservation of mass, the balance of linear momentum, the balance of energy, and the Clausius-Duhem inequality:

$$\begin{aligned}
 (4.1) \quad & \rho_t + \operatorname{div}(\rho \mathbf{u}) = 0, \\
 & \rho \frac{D\mathbf{u}}{Dt} = \operatorname{div} \mathbf{T}, \\
 & \rho \frac{D\varepsilon}{Dt} = \mathbf{T} \cdot \mathbf{L} - \operatorname{div} \mathbf{q} + \operatorname{div} \mathbf{w}, \\
 & \rho \theta \frac{D\eta}{Dt} + \operatorname{div} \mathbf{q} + \frac{\mathbf{q} \cdot (\operatorname{grad} \theta)}{\theta} \geq 0,
 \end{aligned}$$

where $\frac{Df}{Dt} = f_t + \mathbf{u} \cdot \nabla f$ and

1. $\rho = \rho(\mathbf{x}, t)$ is the density of the fluid at the point \mathbf{x} at time t ,
2. $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ is the velocity of fluid,
3. $\theta = \theta(\mathbf{x}, t) (> 0)$ is the absolute temperature,
4. $\varepsilon = \varepsilon(\mathbf{x}, t)$ is the specific internal energy per unit mass,
5. $\eta = \eta(\mathbf{x}, t)$ is the specific entropy per unit mass,
6. $\mathbf{T} = \mathbf{T}(\mathbf{x}, t)$ is the Cauchy stress tensor.
7. $\mathbf{q} = \mathbf{q}(\mathbf{x}, t)$ is the heat flux vector,
8. $\mathbf{L} = \operatorname{grad} \mathbf{u}$.

The main difference with the classical thermodynamics is the $\operatorname{div} \mathbf{w}$ term and \mathbf{w} is called the interstitial work flux representing spacial interactions of longer range. They have proved that for a given Helmholtz free energy $\psi(\rho, \theta, \mathbf{d})$, the following forms of \mathbf{w} and \mathbf{T}

$$\begin{aligned}
 (4.2) \quad & \mathbf{w} = \rho \dot{\psi}_{\mathbf{d}} + \bar{\mathbf{w}}, \\
 & \mathbf{T} = (-\rho^2 \psi_{\rho} + \rho \mathbf{d} \cdot \psi_{\mathbf{d}} + \rho^2 \nabla \cdot \psi_{\mathbf{d}}) \mathbf{I} - \rho \mathbf{d} \otimes \psi_{\mathbf{d}}
 \end{aligned}$$

are compatible with (4.1d). Here, $\rho^2 \psi_{\rho}(\rho, \theta, 0)$ is the pressure and $\bar{\mathbf{w}}$ is the "static" portion of the interstitial work flux \mathbf{w} . They have shown that if the material possesses a center of symmetry, $\bar{\mathbf{w}} = 0$. In what follows, we consider the materials which possess the center of symmetry. They also have observed that the classical forms of viscosity and conductivity tensors are compatible.

In this note we state a result concerning the existence of a unique local smooth solution in the two-dimensional isothermal motion of the Korteweg type materials where the viscous effect is also included. The 3-dimensional case can be discussed similarly. In what follows, we state the assumptions on the Helmholtz free energy and derive the system that we shall discuss. We assume that the Helmholtz free energy is given by

$$(4.3) \quad \psi = F(\rho) + \frac{\nu}{2\rho}(\rho_x^2 + \rho_y^2),$$

where F is a smooth function of ρ and ν is a positive constant. This choice is to make the terms appearing in (4.4) as simple as possible, yet to reflect the effect of the higher order terms of ρ .¹

With the choice the Helmholtz free energy given in (4.3) and with $\lambda = -\frac{1}{3}\mu$, the system then becomes

$$(4.4) \quad \begin{aligned} \rho_t + (\rho u)_x + (\rho v)_y &= 0, \\ (\rho u)_t + (\rho u^2)_x + (\rho uv)_y &= (T_{11})_x + (T_{12})_y, \\ (\rho v)_t + (\rho uv)_x + (\rho v^2)_y &= (T_{21})_x + (T_{22})_y, \end{aligned}$$

where u and v are the x and y component of velocity and

$$(4.5) \quad \begin{aligned} \mathbf{T} &= \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix} \\ &= \left\{ -p + \frac{\nu}{2}(\rho_x^2 + \rho_y^2) + \nu\rho\Delta\rho \right\} \mathbf{I} - \nu \begin{pmatrix} \rho_x^2 & \rho_x\rho_y \\ \rho_x\rho_y & \rho_y^2 \end{pmatrix} + \mathbf{V}, \end{aligned}$$

$$(4.6) \quad p = \rho^2 F'(\rho),$$

and

$$(4.7) \quad \begin{aligned} \mathbf{V} &= \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \\ &= \mu \{ (\text{grad} u) + (\text{grad} u)^T - \frac{2}{3}(\text{div} u)\mathbf{I} \}. \end{aligned}$$

Here, \mathbf{I} is the unit rank-two tensor and a superscript T denotes the transpose of a tensor. Since we discuss the existence of a local solution, we do not need the monotonicity of the pressure on ρ . Further computation simplifies the $\text{div} \mathbf{T}$ term

$$(4.8) \quad \text{div} \mathbf{T} = -\nabla p + \nu\rho\nabla(\Delta\rho) + \text{div} \mathbf{V}.$$

We discuss the local existence for the pure initial value problem of (4.4) with the initial data given by

$$(4.9) \quad (\rho, u, v)(x, y, 0) = (\rho_o, u_o, v_o)(x, y).$$

¹Another reasonable choice is to change the last term in (4.3) with $\frac{\nu}{2}(\rho_x^2 + \rho_y^2)$. Although this choice may be physically more realistic, mathematically it is more cumbersome to handle. For example, the expression for $\text{div} \mathbf{T}$ is very complicated. Therefore, we do not discuss this case (See (4.8)).

We assume that the initial data satisfy

$$(4.10) \quad (\rho_0 - \bar{\rho}_0, u_0, v_0) \in H^k(R^2), \quad \bar{\rho}_0 \geq \delta > 0,$$

where $k \geq 4$ and $\bar{\rho}_0 > 0$ is a positive constant. Denote by $\|\cdot\| \equiv \|\cdot\|_0$ the L^2 norm and by $\|\cdot\|_k$ the k -th order Sobolev norm. Set

$$\|w\|_{0,T}^2 \equiv \sup_{0 \leq t \leq T} (\|w(t)\|^2 + \|\nabla \rho(t)\|^2) + \int_0^T (\|\nabla u(t)\|^2 + \|\nabla v(t)\|^2) dt$$

and

$$\|w\|_k^2 = \sum_{|j| \leq k} \|\partial_{x,y}^j w\|^2,$$

where $w \equiv (\rho, u, v)$. The main result is stated as follows.

Theorem 4.1 *For any initial data (ρ_0, u_0, v_0) such that $\rho_0 \geq \delta > 0$ and $(\rho_0 - \bar{\rho}_0, u_0, v_0) \in H^k(R^2)$ ($k \geq 4$) where $\bar{\rho}_0 > 0$ is a constant, there exists a $T > 0$ such that in $t \in [0, T]$, the Cauchy problem (4.4), (4.9) has a unique solution (ρ, u, v) such that $\rho - \bar{\rho}_0 \in L^\infty([0, T]; H^{k+1}(R^2))$ and $(u, v) \in L^\infty([0, T]; H^k(R^2))$ and*

$$\|w\|_k^2 \leq C_k \|w_0\|_k^2 + \|\rho_0\|_{k+1}^2.$$

Since the linearized problem of (4.4), (4.9) is not of any classical type, the existence of solutions is not known even for the linearized problem. We prove the existence of solutions for the linearized problem by establishing an energy estimate for the dual problem and then using the dual argument.

References

- [1] Andrews, G. and J.M. Ball, Asymptotic behaviour and change of phase in one-dimensional nonlinear viscoelasticity, J. Diff. Eqns. 44 (1982), 306-341.
- [2] Bronsard, L. and R.V. Kohn, On the slowness of phase boundary motion in one space dimension, Comm. Pure Appl. Math. 43 (1990), 984-997.
- [3] Carr, J., M.E. Gurtin, and M. Slemrod, One dimensional structured phase transitions under prescribed loads, J. Elasticity 15 (1985), 133-142.
- [4] Carr, J., M.E. Gurtin, and M. Slemrod, Structured phase transitions on a finite interval, Arch. Rat. Mec. Anal. 86 (1984), 317-351.
- [5] Carr, J. and R.L. Pego, Metastable patterns in solutions of $u_t = \epsilon^2 u_{xx} - f(u)$, Comm. Pure Appl. Math. 42 (1989), 523-576.
- [6] Dunn, J.E. and J. Serrin, On the thermodynamics of interstitial working, Arch. Rat. Mech. Anal. 88 (1985), 95-133.

- [7] Fusco, G. and J.K. Hale, Slow-motion manifolds, dormant instability, and singular perturbations, *J. Dyn. Diff. Eqns.* 1 (1989), 75-94.
- [8] Grant, C.P., Slow motion in one-dimensional Cahn-Morral systems, preprint CDSNS92-78, Georgia Institute of Technology.
- [9] Hagan, R. and M. Slemrod, The viscosity-capillarity admissibility criterion for shocks and phase transitions, *Arch. Rat. Mech. Anal.* 83 (1984), 333-361.
- [10] Hattori, H. and K. Mischaikow, A dynamical systems approach to a phase transition problem, *J. Diff. Eqns.* 94 (1991), 340-378.
- [11] Hattori, H. and K. Mischaikow, On the slow motions of phase boundaries in the Korteweg theory of capillarity, to appear in *Dynamic Systems and Applications*.
- [12] Korteweg, D.J., Sur la forme que prennent les équations des mouvement des fluides si l'on tient compte des forces capillaires par des variations de densité, *Arch. Neerl. Sci. Exactes. Nat. Ser. II* 6 (1901), 1- 24.
- [13] Nirenberg, L., On elliptic partial differential equations, *Annali della Scuola Norm. Sup.-Pisa* 13 (1959), 115-162.
- [14] Pego, L.R., Phase transitions in one-dimensional nonlinear viscoelasticity: Admissibility and stability, *Arch. Rat. Mech. Anal.* 97 (1987), 353-394.
- [15] Serrin, J., Phase transition and interfacial layers for van der Waals fluids, in "Proceedings of SAFA IV Conference, Recent Methods in Nonlinear Analysis and Applications, Naples, 1980" (A. Camfora, S. Rionero, C. Sbordon, C. Trombetti, Eds.)
- [16] Serrin, J., The form of interfacial surfaces in Korteweg's theory of phase equilibria, *Quart. J. Appl. Math.* 41 (1983), 357-364.
- [17] Slemrod, M., Admissibility criteria for propagating phase boundaries in a van der Waals fluid, *Arch. Rat. Mech. Anal.* 81 (1983), 301-315.
- [18] Slemrod, M., Dynamic phase transitions in a van der Walls fluid, *J. Diff. Eqns.* 52 (1984), 1-23.

Singular Value Computation on a Fat-Tree Network

Tong J. Lee

School of Electrical Engineering
Cornell University
Ithaca, New York 14853

Franklin T. Luk

Department of Computer Science
Rensselaer Polytechnic Institute
Troy, New York 12180

Daniel L. Boley

Department of Computer Science
University of Minnesota
Minneapolis, Minnesota 55455

Abstract

The Singular Value Decomposition (SVD) is a matrix tool that plays a critical role in many applications: for example, in signal processing, it is often necessary to calculate the SVD in real time. We present here a new technique for computing the SVD on a parallel architecture whose processors are connected via a fat-tree. We tested our idea on the Connection Machine CM-5, and achieved efficiency up to 40% even for moderately sized matrices.

KEYWORDS. Singular value decomposition, parallel Jacobi algorithm, fat-tree, CM-5.

1 Introduction

Let A be a real $m \times n$ matrix. Its singular value decomposition (SVD) is given by

$$A = U \Sigma V^T,$$

where U and V are respectively $m \times m$ and $n \times n$ orthogonal matrices and Σ is an $m \times n$ diagonal matrix. The best approach to parallel SVD computation is apparently one of the Jacobi type; see, e.g., [1], [2], [4], [5], [7], [11], [12]. In this paper, we will discuss the efficient implementation of a Jacobi method on a parallel computer with a fat-tree interconnection network. We will propose a new Jacobi ordering for a fat-tree and analyze its behavior both theoretically and experimentally (on a Connection Machine CM-5).

This paper is organized as follows. In the next two subsections, we present the fat-tree architecture and Jacobi algorithm. Section 2 introduces a new fat-tree ordering, and provides some kernel programs. We analyze communication costs on a fat-tree network in Section 3, and discuss implementation results on the CM-5 in Section 4.

1.1 Fat-Tree Architecture

The fat-tree was introduced by Leiserson [10] as a novel approach to interconnect the processors of a general-purpose parallel supercomputer. This communication structure can also be seen in the distributed computing environment, such as a network of workstations.

The routing network of the Connection Machine CM-5 [14] is based on the fat-tree. This parallel machine consists of up to 544 ($= 512 + 32$) nodes for the model at the Army High Performance Computing Research Center (AHPARC) at the University of Minnesota, and 32 nodes at the Northeast Parallel Architectures Center (NPAC) at Syracuse University. Each node of the CM-5 is a SPARC chip which runs at 32 MHz and delivers 22 Mips and 5 Mflops. There is a 64 Kbyte instruction and data cache and a 16 Mbyte memory in each node. All the nodes are synchronized. In October of 1992, two vector units will be installed in each processing node; each vector unit is capable of 64 Mflops peak and 40 Mflops sustained [9]. The control and data networks are connected via a *skinny* fat-tree structure. By *skinny*, we mean that the bandwidth does not increase proportionately to the number of nodes; in particular, the bandwidth is 20Mbyte/sec per node in a group of four processors, 10 Mbyte/sec per node in a group of sixteen, and 5Mbyte/sec overall. So data contention may severely degrade performance when all nodes need to access a large set of data from other nodes through the top level of the tree.

1.2 Jacobi Algorithm

The one-sided Jacobi method [8] generates an orthogonal matrix V such that the columns of the matrix W , given by $W = AV$, are mutually orthogonal. The matrix V can be generated by a sequence of plane rotations $V^{(1)}, V^{(2)}, \dots$, where each $V^{(k)}$ is an identity matrix except for four entries: $v_{ii}^{(k)} = \cos \theta$, $v_{ij}^{(k)} = -\sin \theta$, $v_{ji}^{(k)} = \sin \theta$ and $v_{jj}^{(k)} = \cos \theta$, where (i, j) represents the index pair of the columns of A that $V^{(k)}$ orthogonalizes. The SVD computation requires $O(mn^2)$ operations for an $m \times n$ matrix A . For a limited number of processors, i.e., up to $n/2$ processors, an efficient way is to configure them as a linear array along the horizontal dimension. Columns can be distributed either in blocks or in a wraparound fashion. Note from the above derivation that

each column-pair can be orthogonalized independently, so that we may transform up to p pairs concurrently, where p denotes the number of processors. This method was used for computing the SVD on special machines, e.g., parallel computers such as the Illiac IV [11] and vector processors such as the CYBER 205 [3]. The one-sided Jacobi method is composed of these major steps:

1. Compute the norm of each column.
2. Compute plane rotations to orthogonalize paired columns.
3. Apply the plane rotations to update the columns and the column norms.
4. Permute the columns in a pre-chosen order to generate the next column pairs, and repeat the process from step 2.

If the column pairs are distributed to different processors, then step 4 requires communication. In the case of a two-dimensional mesh (as in the ILLIAC IV), each column is itself distributed among different processors and step 3 requires that the rotation parameters be transmitted to all the processors containing each given column pair. In the case of a one-dimensional array, each column pair is stored entirely in one processor and significant speedup is possible if vector units are present within each processor.

In this paper, we use the one-dimensional array, with each processor storing two blocks of columns. That is, we use a *block* Jacobi algorithm, in which the column blocks are circulated according to a given ordering to be defined, and the *cyclic-by-rows* ordering [6] is used within each block.

2 New SVD Algorithm

In the past, when the hypercube interconnection topology was in vogue, several Jacobi ordering schemes were proposed [1], [4], [7] to utilize the hypercube structure. Here, for a one-dimensional array of processors with no wraparound, a chess-tournament ordering [2] may be chosen because it does not waste processing power or memory space. However, communication requires a two-way transmission of columns between adjacent processors. An alternative is a ring ordering [4] which uses only one-way transmission, but it requires a wraparound connection. To develop an ideal ordering for a fat-tree, we aim to minimize the total path length by using the extra bandwidth of a fat-tree.

2.1 Fat-Tree Ordering

It is easiest to describe this ordering by an example. In Figure 1 we show the case for sixteen columns and eight processors. For pedagogic reasons, we use a base 8 numbering of the indices and so $A=8$, $B=9$, ..., $H=15$. The XOR (exclusive-or) column is the binary XOR of the column indices: at each step, the XOR value of each index pair is the same, and from one step to the next this quantity follows the Gray code. The *cost-to-this-step* column denotes the maximum number of levels up the tree the messages must travel to reach their destinations from the previous step. In general, if there are p processors and two columns per processor, then a sweep requires $2p - 2$ steps. We save one step per sweep because the last step of sweep i can be included as the first step for sweep $i + 1$.

Step	Ordering of Index Pairs	XOR	Cost to This Step
0.	(01)(23)(45)(67)(AB)(CD)(EF)(GH)	0001	NA
<i>Forward Sweep</i>			
1.	(03)(12)(47)(56)(AD)(BC)(EH)(FG)	0011	1
2.	(02)(13)(46)(57)(AC)(BD)(EG)(FH)	0010	1
3.	(06)(17)(24)(35)(AG)(BH)(CE)(DF)	0110	2
4.	(07)(16)(25)(34)(AH)(BG)(CF)(DE)	0111	1
5.	(05)(14)(27)(36)(AF)(BE)(CH)(DG)	0101	2
6.	(04)(15)(26)(37)(AE)(BF)(CG)(DH)	0100	1
7.	(0E)(1F)(2G)(3H)(4A)(5B)(6C)(7D)	1100	3
8.	(0F)(1E)(2H)(3G)(4B)(5A)(6D)(7C)	1101	1
9.	(0H)(1G)(2F)(3E)(4D)(5C)(6B)(7A)	1111	2
10.	(0G)(1H)(2E)(3F)(4C)(5D)(6A)(7B)	1110	1
11.	(0C)(1D)(2A)(3B)(4G)(5H)(6E)(7F)	1010	3
12.	(0D)(1C)(2B)(3A)(4H)(5G)(6F)(7E)	1011	1
13.	(0B)(1A)(2D)(3C)(4F)(5E)(6H)(7G)	1001	2
14.	(0A)(1B)(2C)(3D)(4E)(5F)(6G)(7H)	1000	1
<i>Backward Sweep</i>			
13.	(0B)(1A)(2D)(3C)(4F)(5E)(6H)(7G)	1001	1
12.	(0D)(1C)(2B)(3A)(4H)(5G)(6F)(7E)	1011	2
11.	(0C)(1D)(2A)(3B)(4G)(5H)(6E)(7F)	1010	1
10.	(0G)(1H)(2E)(3F)(4C)(5D)(6A)(7B)	1110	3
9.	(0H)(1G)(2F)(3E)(4D)(5C)(6B)(7A)	1111	1
8.	(0F)(1E)(2H)(3G)(4B)(5A)(6D)(7C)	1101	2
7.	(0E)(1F)(2G)(3H)(4A)(5B)(6C)(7D)	1100	1
6.	(04)(15)(26)(37)(AE)(BF)(CG)(DH)	0100	3
5.	(05)(14)(27)(36)(AF)(BE)(CH)(DG)	0101	1
4.	(07)(16)(25)(34)(AH)(BG)(CF)(DE)	0111	2
3.	(06)(17)(24)(35)(AG)(BH)(CE)(DF)	0110	1
2.	(02)(13)(46)(57)(AC)(BD)(EG)(FH)	0010	2
1.	(03)(12)(47)(56)(AD)(BC)(EH)(FG)	0011	1
0.	(01)(23)(45)(67)(AB)(CD)(EF)(GH)	0001	1
<i>Forward Sweep</i>			
1.	(03)(12)(47)(56)(AD)(BC)(EH)(FG)	0011	1

Figure 1. Fat-tree Ordering based on the Gray code
(eight processors and sixteen columns).

2.2 Kernel Programs

To see how to write a simple node program to generate the fat-tree ordering, we use the following observations from the example in Figure 1. To simplify the presentation, we consider only the forward sweep. At each step, each processor must communicate with a remote processor whose label differs in one bit. The basis for our kernel presented here is to compute a mask such that the

exclusive-or of the mask with the current processor label yields the remote processor label. When using the Gray code, this mask can be computed using only the step number – it is independent of the processor label.

We also use the following observations. First, we use the fact that the XOR's follow the Gray code. Second, we observe that during the second half of the forward sweep (steps 7-14), the lower half of the columns (numbers 0,...,7 in Figure 1) remain fixed in the processor with the same number. Hence the location of the remaining columns is fixed entirely by the Gray code. Third, we observe that the first half of the steps (steps 0-6) amount to doing a Gray code fat-tree ordering on each half of the processor array separately. The only remaining step is the transition from the first half to the second half (step 6 to step 7). Hence we can define the ordering for these steps recursively from the smaller cases.

We can summarize the steps for the forward sweep in the following procedure, in a pseudo-MATLAB notation assuming for the sake of simplicity of the presentation that the sends do not block.

```
% Node program for processor ProcNo for one forward sweep using an array of
% NProcs processors. Assume Column(1) and Column(2) are the head and tail
% columns, respectively, in the local memory.
```

```
Orthogonalize_Individual_Column_Blocks % (within each block);
```

```
for StepNo = 1:2*NProcs-2,
```

```
    Pairwise_Orthogonalize_Column_Blocks;
```

```
    %% for each processor, figure where the data goes to and send it.
```

```
    [Mask,ColumnSwitch] = MakeMask(StepNo,ProcNo,NProcs);
    RemoteProcNo = XOR(ProcNo,Mask);
```

```
    Send Column(2) to remote processor RemoteProcNo;
```

```
    if ColumnSwitch == rotate,
```

```
        Column(2) = Column(1);
```

```
        Column(1) = receive_from(RemoteProcNo);
```

```
    else
```

```
        Column(2) = receive_from(RemoteProcNo);
```

```
    end;
```

```
end;
```

```

function [Mask,ColumnSwitch]=MakeMask(StepNo,ProcNo,NoProcs);

% ColumnSwitch indicates which column of the pair is to be sent/received.

% Mask is the XOR Mask so that RemoteProcNo = XOR(ProcNo,Mask).
% The Mask is computed independent of the processor label ProcNo.

% Handle first 2 steps as special cases to start recursion
if StepNo <= 2,
    Mask=1;
    ColumnSwitch = tail;
    if rem(ProcNo,2) == 1 & StepNo == 1, ColumnSwitch = rotate; end;

% First half of sweep: pretend this is a separate fat tree sweep on each
% half of the processor array.
else if StepNo < NoProcs-1,
    [Mask,ColumnSwitch] = MakeMask(StepNo,rem(ProcNo,NoProcs/2),NoProcs/2);

% Middle of sweep: here is first exchange through top of tree.
else if StepNo == NoProcs-1,
    Mask = NoProcs/2;
    ColumnSwitch = tail;
    if ProcNo >= NoProcs/2, ColumnSwitch = rotate; end;

% Last half of sweep: only tail columns move, figure Mask using Gray codes.
else if StepNo > NoProcs-1,
    Mask = xor(gray(StepNo),gray(StepNo+1));
    ColumnSwitch = tail;

end;

```

2.3 Test of Convergence

For a fat-tree ordering, any consecutive $2p - 2$ (or even $2p - 1$) steps may not constitute one sweep. We must complete a sweep, either forward or backward, to ensure that all column pairs have been orthogonalized. The convergence test is simple. We maintain a one-bit counter in every processor. The counter is reset at the beginning of every sweep, and is set whenever a column pair needs to be orthogonalized. At the end of the sweep, a global or operation is performed and convergence is achieved if no bit has been set.

3 Analysis on a Binary Fat-Tree Network

We consider a binary fat-tree with p processors, and assume that the communication time from one processor to another is determined by the number of links a message has to traverse and the capacity of these links. Our assumption is supported by experimental results reported in [13]. Define a channel to be the communication link between any two adjacent nodes: here a node can be a processor or an internal switching element. The capacity of a channel equals the number of parallel wires in the channel, and thus the maximum number of simultaneous bit-serial messages it can support [10]. Denote the capacity of the channels at the bottom level by γ . Label the levels from bottom up as level 1, 2, ..., so that the capacity of the channels at level l is given by $2^{l-1}\gamma$. Let us ignore start-up and latency costs. Within a single problem, all the messages have the same size and thus we measure the cost of multiple message transmission using *path length*.

For the ring ordering, at each step a message always goes through the top level and the maximum path length equals $2\log p$ (unless otherwise stated, we use base 2 logarithms). Since there is at most one message at each channel, congestion never occurs and it takes $2p - 1$ steps to complete one sweep. The total path length equals $(4p - 2)\log p$.

The fat-tree ordering does not cause congestion on a fat-tree network. Hence it suffices to count the number of times that each level is used. Denote that count by $c(p, l)$. Consider the forward sweep. We see from Figure 1 that with $p = 8$ processors, the top level is used in two transition steps, the middle level in six steps and the bottom in fourteen steps. The first six steps correspond to the fat-tree ordering for the first four processors, and also for the second four processors. In the general case of p processors, there are $2p - 2$ steps using $\log p$ levels, of which the first $p - 2$ steps amount to the ordering for $p/2$ processors. When the number of processors doubles to $2p$, we add a new top level and the first $2p - 2$ steps correspond exactly to the p processor ordering. There are an extra $2p$ steps, of which two use the new top level, four use the next level (the old top level), eight use the following level, etc. Formally, we get the recurrence

$$c(2p, l) = c(p, l) + 4(p/2^l) \quad \text{for } l = 1, \dots, \log p,$$

starting with $c(p, \log p) = 2$ and $c(p, l) = 0$ for $l > \log p$. Therefore, $c(p, l) = 4p/2^l - 2$, and the total path length is given by

$$2 \sum_{l=1}^{\log p} c(p, l) = 2[(2p - 2) + (p - 2) + \dots + 14 + 6 + 2] = 8p - 4\log p - 8.$$

For a large p , the path length ratio of the two orderings grows like $\log p/2$, a very attractive result for our new ordering.

4 Connection Machine CM-5

Although the CM-5 network is a 4-way tree, the analysis on 2-way trees is applicable. We take a 4-way tree and expand every interior node into a binary tree consisting of that node with two new children each connected to two of the four former children. The number of levels as well as the path length are doubled. However, the CM-5 is *skinny* and the capacity only doubles at every level. Hence it becomes a *skinny* 2-way tree in which the capacity goes up by $\sqrt{2}$ at each level.

To simplify our analysis, we concentrate on the 32-processor model. So $p = 32$ and there are three tree levels because $\lceil \log_4 p \rceil = 3$. The dominating communication cost for the CM-5 is the overhead time that is spent on address calculation, buffer space management, and so on. Let t_{or} and t_{of} represent the cost of such overhead in each step for the ring and the fat-tree ordering, respectively. Let t_{cf} be the overhead cost for resolving contention in the channels of the CM-5 network when applying the fat-tree ordering, and let t_e be the time for traversing an edge in the network. We note that $t_e < t_{cf} < t_{oh}$, where $t_{oh} \in \{t_{or}, t_{of}\}$, $t_{cf} \approx t_{oh}$, and $t_e \in (t_{oh}/10^3, t_{oh}/10^2)$. The overheads t_{or} and t_{of} depend on the data size and are of equal magnitude.

We proceed to compute the coefficient for t_e , which we assume to equal the number of messages that traverse the channels in one sweep. For the ring ordering, there is no congestion in the networks. So the coefficient for t_e is $2 \cdot 63 \cdot 3 (=378)$, and the total cost equals $63 t_{or} + 378 t_e$. For the fat-tree ordering, we observe that level 1 is visited 62 times, level 2 fourteen times, and level 3 two times. We model the resolution of the contention by sending messages in batches. Messages through level 2 must be sent in two batches and messages through level 3 in four batches, in order to avoid contention. Hence we account for the thinness of the CM-5 network by assigning a weight of two to level 2 and a weight of four to level 3. The total path length is $2(62 + 2 \cdot 14 + 4 \cdot 2) = 196$ and the total cost equals $62 t_{of} + 196 t_e + t_{cf}$. Thus, on the CM-5 the fat-tree ordering may not outperform the ring ordering because of the extra cost associated with message contention.

4.1 Experimental Results

In Table 1 we present implementation results on a 32-node CM-5 for random $n \times n$ matrices with n ranging from 64 to 1024. The program was written in Fortran and each experiment repeated ten times. We measured the overall and computation (by disabling communication) costs for one sweep, and estimated the communication cost by subtracting the latter from the former. Our results show that, despite the message congestion that it causes on the CM-5, the fat-tree ordering gets more competitive as n grows, justifying our effort to minimize the total message path length (see also [13]). The mflops (million floating-point operations per second) figures in Table 2 are computed based on the count that $8n^3$ flops are required for one sweep. We conjecture that the *compute* performance deteriorates when n gets beyond 512 because the cache is no longer large enough to hold the huge column blocks. Nonetheless, our implementation results shows how, as the message size increases (hence t_e increases [13]), the fat-tree ordering quickly becomes competitive.

	n	64	128	256	512	1024
Overall	Ring	$7.595 e^{-2}$	$3.229 e^{-1}$	2.628	$1.794 e^1$	$1.380 e^2$
	Fat-tree	$8.134 e^{-2}$	$3.481 e^{-1}$	2.237	$1.795 e^1$	$1.361 e^2$
Compute	Ring	$3.013 e^{-2}$	$2.320 e^{-1}$	1.871	$1.493 e^1$	$1.309 e^2$
	Fat-tree	$3.436 e^{-2}$	$2.420 e^{-1}$	1.878	$1.493 e^1$	$1.310 e^2$
Communicate	Ring	$4.582 e^{-2}$	$0.909 e^{-1}$	0.757	3.010	7.110
	Fat-tree	$4.698 e^{-2}$	$1.061 e^{-1}$	0.359	3.020	5.140

Table 1. CPU Time (seconds) of Ring and Fat-Tree Orderings

	n	64	128	256	512	1024
Overall	Ring	27.61	51.96	51.07	59.85	62.25
	Fat-tree	25.78	48.20	60.00	59.82	63.11
Compute	Ring	69.60	72.32	71.74	71.92	65.62
	Fat-tree	61.03	69.33	71.47	71.92	65.57

Table 2. Mflops Rates of Ring and Fat-Tree Orderings

Acknowledgements

The work of F. T. Luk was supported in part by start-up funds at the Rensselaer Polytechnic Institute. The authors thank the AHPARC and NPAC for time on the CM-5, and Richard Brent and Lennart Johnsson for valuable discussions on CM-5 communication and hardware issues.

References

- [1] C. H. BISCHOF, *The two-sided block Jacobi method on a hypercube*, in Hypercube Multiprocessors, M. T. Heath, ed., SIAM, 1988, pp. 612-618.
- [2] R. P. BRENT AND F. T. LUK, *The solution of singular-value and symmetric eigenvalue problems on multiprocessor arrays*, SIAM J. Sci. Statist. Comput., 6 (1985), pp. 69-84.
- [3] P. P. M. DE RIJK, *A one-sided Jacobi algorithm for computing the singular value decomposition on a vector computer*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 359-371.
- [4] P. J. EBERLEIN AND H. PARK, *Efficient implementation of Jacobi algorithms and Jacobi sets on distributed memory architectures*, J. Par. Distrib. Comput., 8 (1990), pp. 358-366.
- [5] L. M. EWERBRING AND F. T. LUK, *Computing the singular value decomposition on the Connection Machine*, IEEE Trans. Computers, 39 (1990), pp. 152-155.
- [6] G. E. FORSYTHE AND P. HENRICI, *The cyclic Jacobi method for computing the principal values of a complex matrix*, Trans. Amer. Math. Soc., 94 (1960), pp. 1-23.
- [7] G. R. GAO AND S. J. THOMAS, *An optimal parallel Jacobi-like solution method for the singular value decomposition*, in Internat. Conf. Parallel Proc., 1988, pp. 47-53.
- [8] M. R. HESTENES, *Inversion of matrices by biorthogonalization and related results*, J. Soc. Indust. Appl. Math., 6 (1958), pp. 51-90.

- [9] S. L. JOHANSSON. Private communication, September 1992.
- [10] C. E. LEISERSON, *Fat-trees: Universal networks for hardware-efficient supercomputing*, IEEE Trans. Computers, c-34 (1985), pp. 892-901.
- [11] F. T. LUK, *Computing the singular-value decomposition on the ILLIAC IV*, ACM Trans. Math. Softw., 6 (1980), pp. 524-539.
- [12] ———, *A triangular processor array for computing singular values*, Lin. Alg. Applics., 77 (1986), pp. 259-273.
- [13] R. PONNUSAMY, A. CHOUDHARY, AND G. FOX, *Communication overhead on CM5: an experimental performance evaluation*, in Frontier 92, Fourth Symp. on the Frontiers of Massively Parallel Computation, IEEE, 1992, pp. 108-115.
- [14] THINKING MACHINES CORPORATION, *The Connection Machine CM-5 Technical Summary*, October 1991.

A Maximal Invariant Framework for Adaptive Detection

Sandip Bose, Allan Steinhardt

Department of Electrical Engineering, Room 324 ETC

Cornell University, Ithaca, NY, 14853-5401

abstract

We introduce a framework for exploring array detection problems in a reduced dimensional space by exploiting the theory of invariance in hypothesis testing. This involves calculating a low dimensional basis set of functions called the maximal invariant, the statistics of which are often tractable to obtain, thereby making analysis feasible and facilitating the search for tests with some optimality property. Using this approach, we obtain a locally most powerful test for the unstructured covariance case and show that the Kelly and AMF detectors form an algebraic span for any invariant detector. Applying the same framework to structured covariance matrices, we gain some insights and propose several new detectors which are shown to perform as well or better than existing detectors.¹

Introduction

The problem of detecting a signal vector of known direction but unknown strength in Gaussian noise whose covariance matrix is unknown has received much attention lately. In [7], Reed et al used the sample covariance estimate from secondary (signal free) data vectors to derive a weight vector for use in an adaptive matched filter (AMF) detector. Kelly^[3] used the Generalized Likelihood Ratio (GLR) procedure to derive a constant false alarm rate (CFAR) test. Both methods assume that the covariance matrix is completely unknown (unstructured). In many applications, however, the array geometry and partial information of the noise environment (number of interferers, rough bearing estimates etc.) impose a structure on the covariance matrix. It has been shown in [1] and [4] that the use of structured covariance estimates results in a significant improvement in performance in terms of gain in PD and reduction in the number of secondary data vectors required.

In this research, we introduce a framework for studying the optimality properties of these tests. We consider the following structure for the covariance matrix:

$$R = \Psi B \Psi^H + \lambda R_0 \quad (1)$$

where $R(N \times N)$ is the covariance matrix, $\Psi(N \times d)$ spans a rank- d subspace and R_0 is a known covariance matrix. For this research, we assume that Ψ is known while B and λ are not. This structure not only corresponds to the case of a low rank interference component in a dominant subspace (which frequently arises in narrow-band processing when the noise has an interference component due to a small number of sources superimposed on the receiver noise which is usually white); but also as a special case reduces to the unstructured matrix when d equals N . We shall therefore work with this model to obtain general results which can then be applied to specific instances.

¹An earlier version of this material was presented at the IEEE ASSP conference, March 1992, San Francisco, CA

Another case is the block diagonal form for the covariance which may be used to model a non-stationary environment.

Unfortunately, it turns out that for these covariance structures, with the signal bearing and waveform known, it becomes intractable to use the GLR procedure to obtain a test statistic.

Consequently, we approach signal detection from the viewpoint of the general theory of hypotheses testing. We model the signal strength μ as deterministic-unknown. This along with the unknown covariance matrix become the parameters describing the distribution of the observed data vectors. The problem of signal detection becomes one of choosing between two disjoint parameter sets based on the observations. Thus we have the following hypothesis testing problem:

Given

$$X_{N \times L} \sim \mathcal{N}(\mu a e_1^H, R \otimes I) \quad (2)$$

where the columns of X are independent data vectors each normally distributed with covariance R as in (1), a is the signal vector (known), possibly present only in the first column and μ is its strength (unknown).

Test

$$H_0 : \mu = 0$$

versus

$$H_1 : \mu \neq 0$$

Note that the covariance is a nuisance parameter which should not affect the decision statistic. This motivates us to reduce the problem as follows. Transformations on the data that induce transformations on the parameters to which the parameter sets are invariant leave the decision problem unchanged. Therefore, the decision statistic should also be invariant to all such transformations. More concretely, this can be formulated as follows:

Let X be the data characterized by the probability distribution P_θ , $\theta \in \Omega$ and let g be a 1 : 1 onto transformation on the sample space such that gX is distributed as $P_{\theta'}$, $\theta' \in \Omega$. This transformation thereby induces a transformation \bar{g} on the parameter space. It is shown in [5] that the set of all transformations g , such that the corresponding induced transformation \bar{g} is a 1 : 1 map of Ω onto itself, form a group.

The decision problem $H_0 : \theta \in \Omega_0$ vs $H_1 : \theta \in \Omega_1$ is invariant to the group of transformations, G , if $\bar{g}\Omega_i = \Omega_i$, $i = 0, 1$ for all $g \in G$. In that case we require the decision statistic to be invariant to all transformations in G .

This principle of invariance (Lehmann^[5]) greatly reduces the class of detectors to be considered and frequently, it may become possible to find a uniformly most powerful test within this smaller invariant class (UMPI), even though no general UMP test may exist. Often, the GLR procedure leads to such a test. In our case, since the GLR is unavailable, we proceed by deriving the group of transformations that leave the problem invariant. From this, we obtain the maximal invariant, which is the algebraic basis for the largest set of independent functions of the data that are invariant to the transformations. These functions separate the sample space into orbits or invariant subsets. Thus, $M(X)$ is a maximal invariant iff

$$M(X) = M(g(X)), \forall g \in G \quad (3)$$

$$M(X_1) = M(X_2) \Rightarrow X_1 = g(X_2) \text{ for some } g \in G$$

It is shown in [5] that all invariant test statistics are functions of the maximal invariant, whose distribution depends on a reduced parameter set (this may eliminate the nuisance parameters from the problem, which is a very desirable feature). The maximal invariant turns out to be a small set and it is feasible to come up with a reasonable test statistic.

A Maximal Invariant Framework

To begin, consider as a special case of (1) the following structure

$$R = \begin{pmatrix} R_\psi & \mathbf{0} \\ \mathbf{0} & \sigma^2 I_{(N-d)L} \end{pmatrix} \quad (4)$$

This is completely equivalent to equation (1) with $R_\psi = B + \sigma^2 I_d$, since it can be obtained by a known linear transformation on the data. Again, for the same reason, we can assume the following form for the signal vector, a , and partition the data matrix accordingly:

$$a = \begin{bmatrix} a_1 \\ \mathbf{0} \\ a_2 \\ \mathbf{0} \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{bmatrix} \quad (5)$$

where x_{11} is 1×1 , x_{12} is $1 \times (L-1)$, x_{21} is $(d-1) \times 1$, x_{31} is 1×1 and x_{41} is $(N-d) \times 1$

We can represent this matrix as a length- NL vector

$$x = \begin{bmatrix} x_{11} \\ x_{21} \\ \text{vec} \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix} \\ x_{31} \\ x_{41} \\ \text{vec} \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix} \end{bmatrix} \quad (6)$$

where $\text{vec}(A) = [A_1^H \ A_2^H \ \dots \ A_N^H]^H$ for $A = [A_1 \ A_2 \ \dots \ A_N]$

This is distributed as a Gaussian vector with mean μa and covariance $\text{diag}(R_\psi \otimes I_L \ \sigma^2 I_{(N-d)L})$ where \otimes denotes the Kronecker product.

This decision problem is invariant to all transformations which preserve the Gaussian nature of the distribution, the mean vector to a scale factor and the structure of the covariance matrix. The largest group of such linear transformations are given by $T(x) = Gx$ where

$$G = \alpha \begin{pmatrix} \begin{pmatrix} 1 & \beta^H \\ 0 & \Gamma \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & U_1 \end{pmatrix} & \mathbf{0} \\ \mathbf{0} & \begin{pmatrix} 1 & 0 \\ 0 & U_2 \end{pmatrix} \end{pmatrix} \quad (7)$$

where $U_1(L-1)$ and $U_2((N-d)L-1)$ are unitary matrices, β^H is $(1 \times L-1)$ and Γ is $(N-1 \times L-1)$.

We show in [8] that the maximal invariant to this group of transformations is given by

$$\begin{aligned} m_1 &= \frac{\|x_{11} - x_{12}x_{22}^H(x_{22}x_{22}^H)^{-1}x_{21}\|^2}{x_{12}(I - x_{22}^H(x_{22}x_{22}^H)^{-1}x_{22})x_{12}^H} \\ m_2 &= x_{21}^H(x_{22}x_{22}^H)^{-1}x_{21} \end{aligned}$$

$$\begin{aligned}
m_3 &= \frac{\|x_{31}\|^2}{\|x_{32}\|^2 + \|x_{41}\|^2 + \|x_{42}\|_F^2} \\
m_4 &= \frac{x_{11} - x_{12}x_{22}^H(x_{22}x_{22}^H)^{-1}x_{21}}{x_{31}}
\end{aligned} \tag{8}$$

A corresponding maximal invariant in the parameter set is given by

$$\begin{aligned}
\theta_1 &= |\mu|^2 |a_1|^2 (R_\psi)_{11}^{-1} \\
\theta_2 &= \sigma^2 (R_\psi)_{11}^{-1}
\end{aligned} \tag{9}$$

Thus we have greatly reduced the dimensionality of the problem. We obtain the density function for the maximal invariant [8] which is now parameterized by θ_1 and θ_2 above. We show there that no UMP test exists for this problem. Further, since θ_2 is a free parameter even under H_0 , the distribution function of the maximal invariant is not completely specified thereunder and hence an invariant decision statistic will, in general, not have the CFAR property. Approximate CFARness is all one can hope for.

For the unstructured case, the maximal invariant reduces even further, to m_1 and m_2 and the corresponding parameter set to a single parameter θ_1 (which is the SNR). The distribution function under H_0 only depends on the dimension of the data set, and so in this case, any invariant decision rule will be CFAR. Again, in this case no UMP test exists. However, in many applications the performance is critical only for low SNR and a locally most powerful invariant test (LMPI) in the limit of zero SNR is of interest. Since, the parameter space is one dimensional, it becomes feasible to obtain the LMPI test statistic following the theory in [6]. The LMPI decision rule in the limit of θ_0 is given by:

$$\frac{\frac{\delta f_\theta(x)}{\delta \theta} |_{\theta=\theta_0}}{f_{\theta_0}} \underset{H_0}{\overset{H_1}{>}} \tau \tag{10}$$

where f is the density function of x with parameter θ .

In [8] we derive the following density function for m_1 and m_2 :

$$\begin{aligned}
f(m_1, m_2) &= k_1 \frac{(1+m_2)^{-N} m_2^{N-2}}{(1+m_1+m_2)^{L-N}} e^{-\frac{\theta_1}{1+m_1+m_2}} \\
&\quad \sum_{k=0}^{L-N} k_2 \left(\frac{\theta_1 m_1}{(1+m_2)(1+m_1+m_2)} \right)^k
\end{aligned} \tag{11}$$

where $k_1 = \frac{L-1!}{N-2!L-N-1!}$ and $k_2 = \frac{L-N!}{L-N-k!k!}$

Applying the rule in (10), we obtain the following LMPI test:

$$\frac{(L-N)t_K - 1}{(1+m_2)(t_K + 1)} \underset{H_0}{\overset{H_1}{>}} \tau \tag{12}$$

where $t_K = m_1/(1+m_2)$ is the Kelly statistic.

The probability of false alarm for this test is given by

$$PFA = \left(1 - \frac{\tau}{L-N}\right)^{L-1} \left(\frac{L-N}{L-N+1}\right)^{L-N} \tag{13}$$

for $\tau \geq 0$. The detection probability is calculated numerically as a finite sum of simple integrals. Preliminary comparisons with the Kelly statistic indicate a slightly better performance at very low SNR (a gain of 0.1 dB for $N = 4, L = 9, pfa = .1$ at -5 dB SNR) at the expense of a degradation in the higher SNR region (0.3dB loss at 10dB SNR). Further simulations are in progress.

Finally, we note that m_1 is exactly the AMF statistic and the Kelly statistic is $m_1/(1 + m_2)$. Thus these two form an equivalent basis set to m_1 and m_2 . This implies that they form an algebraic basis for all invariant detectors and in searching for viable detectors, it is sufficient to look at compositions of them. It is not necessary to explore alternative ways of projecting down the initial raw high-dimensional data.

Detectors for subspace covariance structures

For the structure in (4), the UMPI test does not exist and the notion of LMPI test is not directly applicable either. However based on considerations of the maximum likelihood estimates of the covariance from the signal free data, we obtain an invariant test which reduces to the Kelly statistic for the unstructured case:

$$\frac{\|a^H \hat{S}^{-1} x\|^2}{a^H \hat{S}^{-1} a} \underset{H_0}{\overset{H_1}{>}} \tau \quad (14)$$

Where

$$\hat{S} = \text{diag} \left(\frac{1}{(L-d)(1+m_2)} \begin{bmatrix} x_{12} \\ x_{22} \end{bmatrix} \begin{bmatrix} x_{12}^H & x_{22}^H \end{bmatrix} \right) \quad (15)$$

$$\frac{\|x_{12}\|^2 + \|x_{22}\|^2 + \|x_{32}\|^2}{(N-d)L-1}$$

with the partitioning of the data and the signal vector as in (5) This is shown to be approximately CFAR and the simulation results in [8] show that it outperforms the Kelly test applied to the data truncated to the span of the interference and signal spaces. In fact, it does nearly as well as the clairvoyant colored noise matched filter whose weights are based on the true covariance and which therefore bounds the best achievable performance.

In some cases, the noise level is known to be of the same order of magnitude in each of the subspaces. This situation may arise for block diagonal covariance matrices modelling certain kinds of non-stationary environments. In this case, a CFAR test is shown to perform almost as well as the test proposed by Kelly in [2].

Conclusion and Comments

Detection in an array environment involves projecting down the multivariate data to a scalar statistic. Since any reasonable statistic must satisfy the invariance criterion, the maximal invariant set specifies all the functions one need consider in devising the detector. Since this set is often small, it is feasible to do analysis and search for a detector with some optimality property with the confidence that the search is over the whole class of reasonable detectors. This framework provides an alternate route to conventional methods like the GLRT for arriving at decision rules and further enables the study of their optimality properties. Thus, for the unstructured covariance case studied by Kelly and others, we show that the Kelly and the AMF statistics form a maximal invariant set. Further, we show that a UMPI does not exist and obtain a LMPI test

around 0 SNR. The Kelly detector is seen to perform nearly as well which is a good endorsement for its use. For the structured covariance case where the GLRT breaks down, we again obtain a small invariant set whose statistics can be analyzed. For this case, the UMP test does not exist. We propose several new tests and show via simulation that they are equivalent or better than existing ones.

References

- [1] D. Fuhrmann and F. Robey. Adaptive detection with structured covariance matrices. *IEEE Trans. on Signal Proc.*, submitted.
- [2] E.J. Kelly. Adaptive detection in non-stationary interference, part 1 and part 2. Technical report, Lincoln Laboratory, Massachusetts Institute of Technology, 1985.
- [3] E.J. Kelly. An adaptive detection algorithm. *IEEE Trans. Aerospace and Electronic Systems*, March 1986.
- [4] I.P. Kirsteins and D.W. Tufts. Rapidly adaptive nulling of interference, 1989.
- [5] E.L. Lehmann. *Testing Statistical Hypotheses*, pages 284–286. Wiley-Interscience, second edition, 1986.
- [6] H. Vincent Poor. *An Introduction to Signal Detection and Estimation*, pages 53–54. Springer-Verlag, 1988.
- [7] I.S. Reed, J.D. Mallet, and L.E. Brennan. Rapid convergence rate in adaptive arrays. *IEEE Trans. Aerospace and Electronic Systems*, November 1974.
- [8] S. Bose and A.O. Steinhardt. Adaptive detection with arrays : Insights using a maximal invariant framework. *IEEE Trans. on Signal Proc.*, To be submitted.

Scalable Software Tools for Adaptive Scientific Computation*

Boleslaw K. Szymanski, Can Özturan, and Joseph E. Flaherty

Department of Computer Science, Rensselaer Polytechnic Institute

Troy, New York 12180-3590, USA

Abstract

With a constant need to solve scientific and engineering problems of ever growing complexity, there is a corresponding need for software tools that assist in generating solutions with minimal user involvement. Parallel computation is becoming indispensable in solving the large-scale problems that arise in science and engineering applications. Adaptivity is at the center of efficient methods for solving partial differential equations often used in such applications. Yet the use of parallel computation and adaptive techniques is limited by the high cost of developing the needed software. To overcome this difficulty, we advocate a comprehensive approach to the development of scalable architecture-independent software for adaptive solutions of partial differential equations.

Our approach is based on program decomposition, parallel computation synthesis and run-time support for adaptive computations. Parallel program decomposition is guided by the source program annotations provided by the user. A family of annotation languages has been designed for this purpose. The synthesis of parallel application is based on configurations that describe overall computation and interaction of its components. Run-time support is responsible for redistributing data and computation during program execution in response to changing computational needs of different subregions during adaptive solution. Adaptive finite difference and finite element procedures tuned to a specific size and type of parallel architecture will be synthesized from components of a decomposed source programs. In this paper, we discuss annotations and configurations suitable for parallel programs written in FORTRAN or in the functional parallel programming language called EPL.

*This work was partially supported by National Science Foundation under grants CCR-8920694 and CDA-8805910

1 Introduction

Several problems listed as "grand challenges" of the Federal High-Performance Computing Program [16] involve the solution of complex multi-dimensional steady and transient partial differential equations. As the mathematical models include more realistic effects, all of these problems exceed the capabilities of current computer systems. We believe, as others do, that computer performance in the needed range of teraflops can be attained only through massive parallelism. However, raw computing power alone is not sufficient to solve a complex problem. We must ensure that (i) adequate mathematical models are used, (ii) reliable numerical methods are employed to approximate these models, (iii) accurate parallel implementations of the methods are executed, (iv) results are within prescribed numerical accuracy, and (v) parallel implementations use the available computational power efficiently.

Adaptivity, with its associated error estimation and shrewd use of computation only in regions where accuracy requirements are not satisfied, provides the needed numerical reliability and efficiency. Adaptive solutions often converge at rates that are much higher than those obtained by conventional methods using a single grid. At the same time, adaptive methods are challenging from the point of view of programming complexity because they use sophisticated data structures, recursion, run-time domain and method selection, etc. Parallelism adds to this challenge because software development for parallel architectures is more complex than for sequential machines due to the increased complexity of assuring parallel program correctness and efficiency. Parallel program correctness requires the results to be independent of the number and speed of the processors. This scalability requirement can be satisfied only if the parallel tasks are independent of each other or properly synchronized when a dependence exists. Synchronization design and verification are the major source of difficulty in assessing parallel program correctness. Different categories of parallel architectures have led to a proliferation of dialects of standard computer languages. Varying parallel programming primitives for different parallel language dialects greatly limits parallel software portability. Clearly, the large efforts required to develop and implement parallel adaptive solution techniques have hampered their widespread application by scientists and engineers. In addition, poor portability of parallel programs has resulted in duplication of effort and has limited the use of developed systems.

The aims of scientific computation are to further understanding of natural phenomena by implementing and executing mathematical models when experiments would be impractical and/or to supplement experiments when direct measurements are not possible. Large-scale computation requires high performance parallel architectures and efficient program implementation to attain acceptable execution times. To facilitate scientific parallel program development, there is a need for software tools that will support **efficiency** as well as

scalability - the same mathematical models and numerical algorithms are often used

in computations with different accuracy and size and executed with a variable number of processors; hence, the cost of the algorithms used in the software tools should increase slowly with the increase in the number of processors used (e.g. the cost function is poly-logarithmic in the number of processors used),

reusability – basic numerical algorithms frequently appear in different models and different computations,

extensibility – interactive development and stepwise refinement of mathematical models describes an implementation of the new models in terms of changes to the old model.

Design methodology of software tools with the above properties is currently a research goal of great importance. Our approach to such design methodology is based on decomposition and scalable synthesis of parallel programs for scientific and engineering computation. The goal is to enable the users to describe high-level features of a parallel computation and to synthesize computation from numerical algorithms, program fragments, and data structures that are separately implemented. Such decomposition and synthesis can support (i) parallel task formulation and allocation, (ii) data distribution, (iii) run-time optimization, and (iv) rapid prototyping of different parallel implementations.

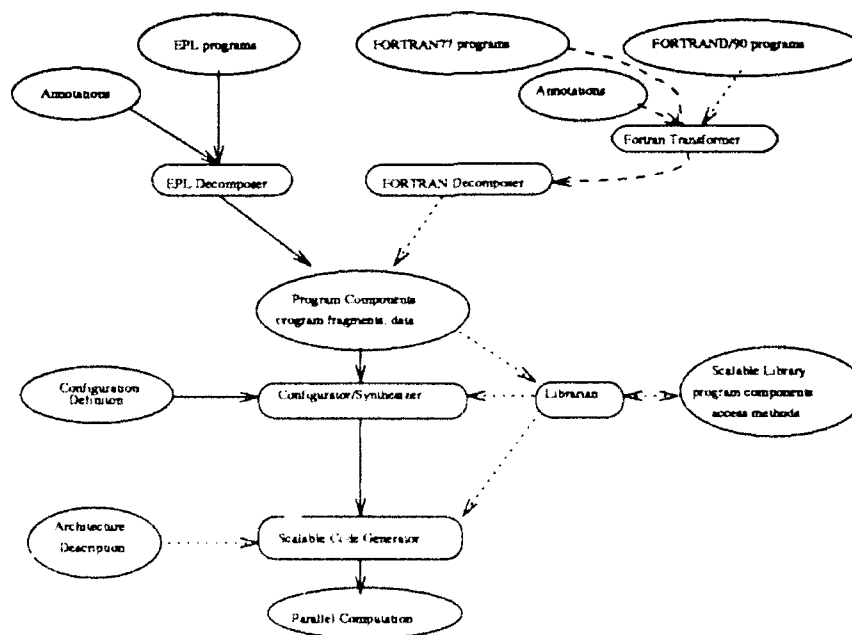


Figure 1: Software tools and their use

The summary view of our approach is given in Figure 1. Program components are created by annotating source programs in FORTRAN or in the functional parallel programming language EPL [14]. FORTRAN programs are transformed into an equational form before decomposition. The configuration definition guides the synthesis of the components into a parallel computation. The synthesized computation together with the architecture description is used by the code generator to produce an object code customized for the target architecture. In the future, we will add a scalable library and an associated librarian to increase versatility of the system. In Figure 1, continuous lines describe implemented paths of the system, broken lines represent paths currently under construction, and dotted lines correspond to paths at an early stage of investigation.

This paper is intended as an overview of the research done towards implementing software tools as envisioned in Figure 1. More technical discussion can be found elsewhere [4, 10, 13, 14, 15].

The paper is organized as follows. Annotations and program decomposition are discussed in Section 2. Program synthesis and the design of the configurator are presented in Section 3. A dynamic load management strategy for adaptive scientific computation on SIMD architectures is a topic of Section 4. Finally, conclusions are outlined in Section 5.

2 Annotations

Annotations provide an efficient way of introducing user's directives for assisting the compiler in parallelization. To be effective, annotations have to be carefully limited to a few constructs. They also should preserve semantics of the original program. In our approach, annotations are introduced solely to limit the allocation of computations to processors. Hence, programs decorated with annotations produce the same results as unannotated program. Consequently, sequential programs that have manifested their correctness over many years of usage are good candidates for parallelization through annotations. By being orthogonal to the program description, annotations support rapid prototyping of different parallel solutions.

2.1 Annotations in EPL

In EPL, each equation can be annotated with the name of the virtual processor on which it is to be executed. Virtual processors can be indexed by the equation's subscripts to identify instances of equations assigned to individual virtual processors. Equation instances annotated by the same virtual processor constitute the smallest granule of parallel computation. An example of the use of EPL annotations in a program for the LU decomposition of a matrix is shown in Figure 2.

```

int: n; /* array size */
real: Ain[*,*],U[*,*],L[*,*];
subscript: i,j;

range.Ain=n; range(2).Ain=n; range.U[j]=j-1; range.L[i]=i;

T[i,j]:A[k,i,j] = if k==1 then Ain[i,j]
                  else if i==Piv[k,k] then A[k-1,Piv[k,k],j]-L[i,k-1]*U[k-1,j];
                  else A[k-1,i,j]-L[i,k-1]*U[k-1,j];
D[j]: L[j,k] = if j==k then 1
               else A[k,j,k]/U[k,k];
D[j]: U[k,j] = A[k,Piv[k,k],j];
D[i]: Piv[k,i] = submax(abs(A[k,i,k]),i:i>=k);

```

Figure 2: LU decomposition of a matrix A in EPL

2.2 Annotations in FORTRAN

As in EPL, the notion of a virtual processor has been introduced in annotations of FORTRAN programs. FORTRAN annotations define blocks of statements associated with a virtual processor, each virtual processor defining a parallel task. Such tasks may include synchronization statements, if they encompass disjoint blocks. FORTRAN virtual processors can have subscripts associated with them to indicate repetition. An example of an annotated FORTRAN segment for the LU decomposition of a matrix is shown Figure 3. The scope of the block extends from the point of definition in the program to the statement labeled 10. In this example, a vector of virtual processors *main*, each associated with a single loop body, is defined. Blocks can also be nested in each other. Such nesting defines a hierarchy of blocks and helps in global program optimization.

Each virtual processor produces data, typically used by other virtual processors, and in turn consumes data produced by others. Performing data-dependence analysis in a style of PTRAN [12], the annotation processor can find the dependencies local to each block and data structures produced and consumed by the block. All data produced by the block are placed in the memory of the corresponding virtual processor. The created parallel tasks are extended by communication statements needed to move data. Parallel tasks associated with virtual processors at the bottom of the block hierarchy are the smallest components used in the program synthesis. An important step towards an efficient parallelization of FORTRAN programs involves an equational transformation during which the equational equivalent of the program is generated. The transformed programs obey the single assignment rule and do not

```

    PARAMETER (N = 50)
    REAL A(N,N), TEMP
    INTEGER IPIV(N)
    DO :: main 10 K = 1, N-1
        IPIV(K) = K
        DO :: pivot 20 L = K+1, N
20          IF (ABS(A(IPIV(K), K)) .LT. ABS(A(L, K))) IPIV(K) = L
        DO :: swap 30 L = K, N
            TEMP = A(K, L)
            A(K, L) = A(IPIV(K), L)
30          A(IPIV(K), L) = TEMP
        DO :: lower 40 L = K+1, N
40          A(L, K) = A(L, K) / A(K, K)
        DO :: up_update 10 L = K+1, N
            DO 10 M = K+1, N
10              A(M, L) = A(M, L) - A(M, K) * A(K, L)
        IPIV(N)=N
    STOP
    END

```

Figure 3: LU Decomposition of a matrix A in FORTRAN

contain any control statements [5]. The transformation is done in the following steps:

Reassignments Elimination: The reassigned variables are replaced by:

- vector (additional dimension) - inside loops,
- variants - in "if" branches and basic blocks.

Condition Analysis: Conditions in the transformed program are analyzed using a Sup-Inf inequality prover [4] and the Kaufl variable elimination method [8] to find pairwise equivalent or exclusive conditions.

Variable's Variants Elimination: Variable variants created in equivalent and exclusive conditions are merged into a single variable.

Additional Dimension Elimination: Memory optimization is performed to replace entire dimensions by windows of few elements for multidimensional variables [15].

The transformed FORTRAN program is then compatible with the programs produced by annotating EPL programs.

2.3 Annotation Processing

Annotation processing includes:

- creating parallel tasks defined by annotated fragments of an original program,
- declaring ports needed to interconnect created tasks into a network,
- building task communication graph that show data dependences between created tasks.

To translate the annotated program into an efficient collection of parallel tasks, it is necessary to embed a spanning tree into the tasks communication graph [11]. The following three criteria are used in selecting such an embedding:

- **Dimension nesting:** If two tasks with different dimensionalities are connected in the task communication graph, the task with more dimensions should be located lower in the spanning tree. If, for example, tasks $T[i][j]$ were located above the tasks $D[j]$ in the spanning tree, the addressing and creation of child tasks in T would involve executing an *if-then* statement in all $i * j$ T tasks.
- **Range nesting:** Whenever possible, tasks sharing the same range should be clustered together in the spanning tree. Variables that share ranges tend to appear in the same equations. Thus, clustering such variables together decreases the number of cross-process references to distributed variables.
- **Data flow:** The total communication cost of the selected spanning tree should be the smallest among all spanning trees satisfying the above two criteria.

Let $G(V,E)$ be a task communication graph with a set of nodes V (representing processors) and a set of edges $E \subseteq V \times V$ representing communication. With each edge $e_{i,j} \in E$ we will associate the cost $c(e_{i,j})$ that represents the volume of data being sent from the processor i to the processor j . With each spanning tree T , we will also associate the distance $d^T(e_{i,j})$ that defines the minimum number of tree edges that have to be traversed on the path from task i to task j . The cost of the spanning tree T can then be defined as:

$$C(T) = \sum_{e_{i,j} \in E} c(e_{i,j}) * d^T(e_{i,j})$$

To minimize the total communication cost we need to find a proper cut-tree, which can be done by solving $|V|$ maximal flow problems. Each maximal flow problem requires $O(|V|^3)$ applications of the Ford-Fulkerson labeling procedure. Hence, finding the solution takes $O(|V|^4)$ steps.

Trees created from annotations of LU decomposition programs are shown in Figure 4 (for EPL and FORTRAN programs).

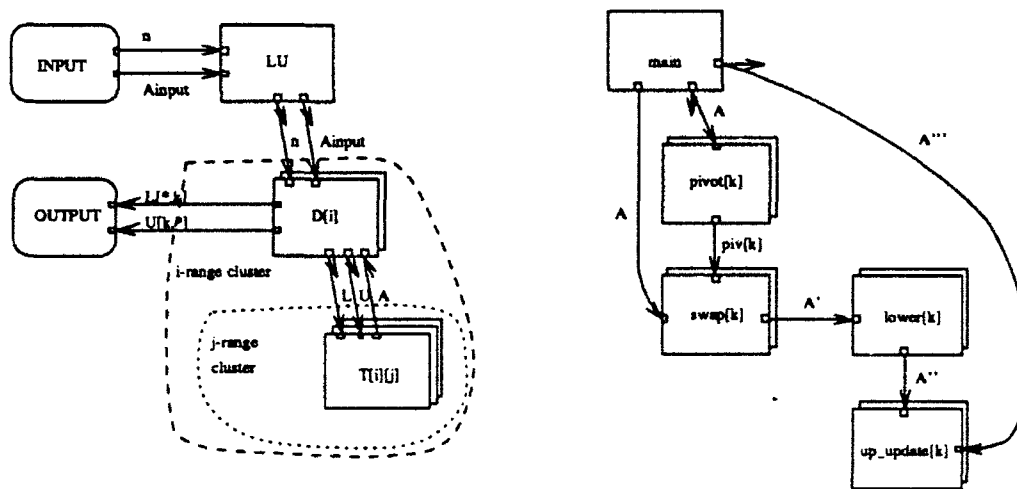


Figure 4: Communication tree for EPL and FORTRAN programs

3 Program Synthesis

In our approach a parallel computation is viewed as a collection of cooperating components. The components are defined during the program decomposition. Their cooperation requires an additional description, called a *configuration*. The configuration guides the process of synthesis. For example, components of the configuration that communicate frequently can be synthesized into a single task. The ratio of physical processors to virtual processors dictates how virtual tasks are to be mapped onto the target architecture. Usually, different annotations result in different configurations and, hence, cause different code to be generated. The user can, therefore, experiment with various annotations to find the one that results in the most efficient code. The configurator uses the dependence graph created during configuration analysis to generate an architecture-independent parallel description which is fed to the code generator.

Configurations define tasks (and their aggregates) and ports. Statements of the configuration represent relations between ports in different tasks. Some of this statements are generated during decomposition (at the subprogram level), others can be supplied by the user (when the programs are integrated into a computation).

Tasks created dynamically can communicate with ports located at parent, child, and sibling tasks (each of those tasks is just a copy of the same program or program fragment, except that a parent task can be arbitrary).

The goal of configuration processing is to establish scheduling constraints for the overall computation. In the parallel computation, individual process correctness is a necessary but not sufficient condition for the correctness of the entire computation. If

a task has input/output ports that belong to a cycle in the configuration graph, then this task's input messages are dependent on the output messages. Such dependences (in addition to dependences imposed by the statements of a task) have to be taken into account in generating the object program for individual tasks; otherwise, loss of messages, process blocking, or even a deadlock can arise.

The algorithm for finding external data dependences has been presented in [13]. It produces *configuration dependence* file used by the synthesizer and the code generator. This file contains a list of the additional, externally imposed data dependences (edges and their dimension types) that need to be added to the task array graph. One task may have several such files, each associated with the different configuration in which this task participates.

4 Run-Time Task Distribution

One of the most challenging problems encountered while implementing adaptive scientific computations on distributed memory machines is run-time mapping of a dynamically changing computational load onto the parallel processors. The published solutions to this problem focus mostly on MIMD architectures and coarse grain parallelism [3]. Recently the following *Rectilinear Partitioning Problem* (RPP) has been considered in [9]: Partition the given $n \times m$ workload matrix into $(N + 1) \times (M + 1)$ rectangles with $N + M$ rectilinear cuts in such a way that the maximum workload among rectangles is minimized. Such optimization is appropriate for adaptive finite element computations on architectures with local communication that is faster than global. Since balanced partitions tend to increase the volume of local vs. global communication, solution to RPP decreases the overall communication costs.

In [10] we investigated adaptive scientific computations on SIMD machines, the problem with similar motivation and applications as RPP [9]. Unlike RPP however, in which the sum of the weights is taken as the cost of a rectangle, we measure the rectangular costs as the ratio of workload to the area of the rectangle that represents the number of processors active in that rectangle. Our approach is motivated by the mesh refinement techniques of the considered adaptive methods and the newly introduced coordinated parallelism on the CM-5 computer. In coordinated parallelism a machine can be partitioned into several parts each running SIMD code. The workload redistribution results in regions that have different time-step and/or grid size; therefore, the same computation is nested in loops with different boundaries. That means that each region either has to be done on the whole machine (sequentially, one after the other on the CM-2) or in a separate partition (in parallel on the CM-5). Each entry in the workload matrix represents the error in the solution obtained by an error estimation procedure [2]. The high-error regions need recomputing to the extent that is proportional to the magnitude of the error. Hence, the number of processors reassigned to each solution region should be proportional to the refinement factor.

Consider a load balancing problem as illustrated in Figure 5 for a one-dimensional

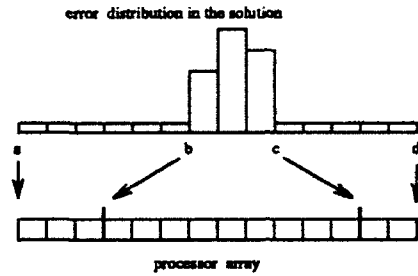


Figure 5: Example of partitioning in one-dimension

problem. The uniform mesh yields the solution with a high error in the interval $b \leq x \leq c$ and within the required accuracy in intervals $a \leq x \leq b$ and $c \leq x \leq d$. Taking the magnitude of an error as an estimate of the work ω_i for each element $i = 1, \dots, n$, we assign a small weight $\epsilon \ll \max_i \{\omega_i\}$ to work estimate in regions $a \leq x \leq b$ and $c \leq x \leq d$. To balance the workload, the majority of the processors should be assigned the interval $b \leq x \leq c$.

In adaptive solutions of partial differential equations parallel tasks perform basically the same computation over different spatial subdomains (intervals for one-dimensional problems) and with different discretization parameter Δx . Let K denote the number of such tasks. It is important to keep this number small for the following reasons. The subdomain interactions are proportional to the number of existing subdomains and in higher dimensions such interactions require time-consuming global communications. In each time step of the subdomain computation, a fraction of executed code is subdomain specific (e.g. in hyperbolic equations the time step has to be set differently in each subdomain). For purely SIMD machines, execution of this code fraction has to be done in K consecutive stages. In each stage, processors in one subdomain are executing while processors belonging to the remaining $K - 1$ subdomains remain idle¹. Therefore, each subdomain associated with a parallel task should represent a localized structure in the solution domain.

Figure 6(a) shows an example of the more difficult two-dimensional case in which a coarse mesh is trivially mapped to the processor mesh. In regions A and B, the mesh must be refined due to the presence of high errors. Hence, we have to spread sub-domains A and B over bigger rectangular sub-sets of processors to improve load balancing as in Figures 6(b) and (c).

If we are employing *mesh-movement* or *static rezone* techniques, the mesh elements are moved into high-error regions. A *global* solution strategy will refine the high-error regions and repeat the entire step of the iteration. Consequently, we will need a re-assignment of processors. A *local* solution strategy, on the other hand, repeats

¹For more general architectures, capable of coordinated parallelism mode of execution (i.e. CM-5), all K subdomains will be able to execute this fraction of code in parallel.

the iteration only where it is needed. Hence, local refinement results in less direct computation and enables more processors to be assigned to regions A and B. However, local refinement requires more interactions between the local and global solutions. Such interactions involve global communication that can outweigh the benefits of an adaptive procedure. Global solutions and mesh-movement techniques require less interactions of this kind. Careful buffering of the high-error regions can increase the number of iterations executed before regridding or mesh movement is needed. This will in turn decrease the frequency of the needed load balancing. It is this global mesh-refinement and mesh movement techniques executed on a mesh connected architectures that motivated us to develop density-type partitioning.

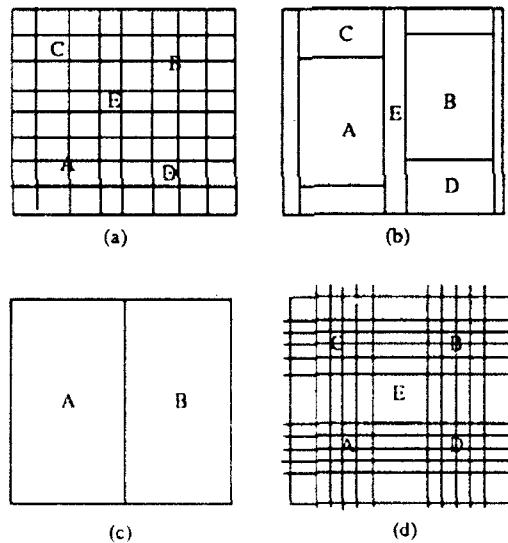


Figure 6: (a) Coarse mesh with high error regions A and B, (b) repartitioning with global refinement (c) repartitioning with local refinement (d) Nicol's partitioning

It should be noted that applying Nicol's [9] partitioning methodology RPP to the example shown in Figure 6(d) results in assigning unnecessary processors to regions C and D. To avoid such waste, we did not restrict our partitioning methodology to rectilinear cuts extending across the whole domain in both dimensions. Instead, in our problem definition and solution [10], we require that K selected rectangles cover the whole domain. The heuristics for the two-dimensional case projects the weights to one-dimension and results in rectilinear cuts extending across the whole dimension in one direction. Figure 6(b) shows an example of this kind of partition.

Let P_K be a set of partitions of a one-dimensional workload array ω_i , $i = 1, \dots, n$ into K subintervals (x_{1_k}, x_{2_k}) , where $1 \leq x_{1_k} \leq x_{2_k} \leq n$, $k = 1, \dots, K$. The one-dimensional workload partitioning problem can be then stated as:

$$\bigoplus \left\{ \bigotimes_k \left\{ \frac{\sum_{i=x_{1_k}}^{x_{2_k}} \omega_i}{f(x_{1_k}, x_{2_k})} \right\} : (x_{1_k}, x_{2_k}) \quad k = 1, \dots, K \in P_K \right\} \quad (1)$$

As shown in Table 1, selecting different meaning for operations \bigoplus and \bigotimes we can obtain different optimization problems from this formulation. For $\bigoplus \equiv \min$, $\bigotimes \equiv \max$ and $f(x_{1_k}, x_{2_k}) = 1$ we obtain the Nicol's problem that has solutions of complexity $O(Kn)$ and $O(n + (K \log n)^2)$ [9].

Problem	\bigoplus	\bigotimes	$f(x_{1_k}, x_{2_k})$	ϵ	e
Nicol's 1D partitioning	\min	\max	1	∞	0
Density-type for PDEs	\min	\max	$(x_{2_k} - x_{1_k} + 1)$	∞	0
Shortest path with k arcs	\min	+	1	∞	0
Density-type for PDEs	\max	\min	$(x_{2_k} - x_{1_k} + 1)$	0	∞

Table 1: *Instances of problem represented by equation (1)*

The problem involving load balancing for adaptive PDE solvers discussed in this section is obtained for $\bigoplus \equiv \min$, $\bigotimes \equiv \max$ and $f(x_{1_k}, x_{2_k}) = (x_{2_k} - x_{1_k} + 1)$, i.e., we divide the sum of the workloads in each partition by the interval length (i.e., the number of processors). There is a similarity between the weighted independent set for interval graphs and our problem [7]. The interval graph for our problem can be created by having a node representing one of the possible subintervals (x_{1_k}, x_{2_k}) with the weight $\sum_{i=x_{1_k}}^{x_{2_k}} \omega_i / f(x_{1_k}, x_{2_k})$ and edges representing the intersections between the subintervals. In such a graph, the independent set of size K which covers the whole interval, $1, \dots, n$, gives the solution to the original problem. We convert that interval graph to a directed acyclic graph (DAG) and apply the shortest path algorithm to find the minimum weight dominating set [10]. This approach results in the optimal algorithm for the one-dimensional case and leads also to a heuristic algorithm that can be easily generalized to two dimensions (by projecting the workloads to one dimension).

5 Conclusion

Our approach is based on the following presumptions:

- Adaptivity is at the center of efficient methods for solving partial differential equations.
- Annotations provide an easy and efficient way for parallelization of existing codes.
- Absence of control statements simplifies program analysis and increases the compiler ability to produce an efficient parallel code.
- Most parallel code optimization problems are NP-hard; hence, development of proper heuristics is important.
- A hierarchical view of parallel computation is helpful in extracting functional parallelism.

Program decomposition through annotations and computation synthesis through configuration can support efficient parallel code generation for domain-specific computation. Adaptivity, with its associated error estimates and shrewd use of computation only in regions where accuracy requirements are not satisfied, can provide the needed numerical reliability and efficiency to parallel computation. Massive parallelism combined with adaptivity offers a promise of true breakthroughs that will allow scientists and engineers to solve the most demanding problems with available resources.

Our research on scalable program synthesis is in its early stages and many issues remain unexplored. Future work on program synthesis should include more work on run-time code optimization. Large applications will measure the efficiency of the generated solutions.

References

- [1] Baber, M.: The Hypertasking Paracompiler - Parallelizing the Game of Life and Other Applications. *Supercomputing Review*. 3, 41-47 (1991)
- [2] Flaherty, J. E., Paslow, P. J., Shephard, M.S. and Vasilakis, J. D., (eds) *Adaptive Methods for Partial Differential Equations*, SIAM, Philadelphia, 1989.
- [3] Berger, M.J., and Bokhari, S.H.: A Partitioning Strategy for Nonuniform Problems on Multiprocessors. *IEEE Trans. on Computers*. C-36, 570-580 (1987)
- [4] Bruno, J., and Szymanski, B.K.: Analyzing Conditional Data Dependencies in an Equational Language Compiler. *Proc. 3rd Supercomputing Conference 1988*, Boston, MA, pp. 358-365. Tampa, FL: Supercomputing Institute 1988
- [5] Ge X., and Prywes, N.S.: Reverse Software Engineering of Concurrent Programs. *Proc. 5th Jerusalem Conference on Information Technology 1990*, Jerusalem, pp. 731-742. Washington, DC: IEEE Computer Science Press 1990

- [6] Gelernter, D., and Carriero, N.: Coordination Languages and their Significance. Comm. ACM. 35, 97-107 (1992)
- [7] Golumbic, M.C.: Algorithmic Graph Theory and Perfect Graphs. New York. NY: Academic Press 1980
- [8] Kauff, T.: Reasoning about Systems of Linear Inequalities. In: Ninth International Conference on Automated Deduction. Aragon. IL, Lecture Notes in Computer Science, pp. 563-72. Berlin-Heidelberg-New York: Springer 1988
- [9] Nicol, D.M.: Rectilinear Partitioning of Irregular Data Parallel Computations. ICASE NASA, Report 91-55, 1991
- [10] Özturan, C., Szymanski, B.K., and Flaherty, J.: Adaptive Methods and Rectangular Partitioning Problem. Proc. Scalable High Performance Computing Conference 1992, Wilmington. VA, pp. 409-415. Washington. DC: IEEE Computer Society Press 1992
- [11] Özturan, C.: Expressing Parallelism in EPL. Rensselaer Polytechnic Institute, Tech. Report No. 90-29, December 1990
- [12] Sarkar, V.: PTRAN - The IBM Parallel Translation System," In: Parallel Functional Languages and Compilers (B.K. Szymanski, ed.). pp. 309-391. New York. NY: ACM Press 1991
- [13] Spier, K., and Szymanski, B.K.: Interprocess Analysis and Optimization in the Equational Language Compiler. In: CONPAR-90. Lecture Notes in Computer Science, pp. 287-98. Berlin-Heidelberg-New York: Springer 1990
- [14] Szymanski, B.K.: EPL - Parallel Programming with Recurrent Equations. In: Parallel Functional Languages and Environments (B.K. Szymanski ed.). pp. 51-104. New York. NY: ACM Press, 1991
- [15] Szymanski, B.K., and Prywes, N.S.: Efficient Handling of Data Structures in Definitional Languages. Science of Computer Programming. 10, pp. 221-245 (1988)
- [16] Walker, T.M. The Federal High Performance Computing Program. Comput. Res. News 1, (1989).

Automated Interpretation of Topographic Maps

T. Cronin
CECOM Signals Warfare Directorate
Warrenton VA 22186-5100

Abstract: Some new results which impact favorably upon the issue of automated topographic map interpretation are presented. Actually, the "new" results consist of heretofore undiscovered applications of two concepts known to mathematicians and computer scientists for many years: binary search, and the normal vector. Binary search is extended by the technique from not only one to two dimensions, but arguably to three dimensions, since topographical maps are two-dimensional representations of three-dimensional surfaces. Such maps are essentially sorted hierarchies of nested contours, which form a multiply-connected subdivision of the plane. The perimeter of a subdivision element is defined by a set of contours of extremal elevation and the edges of the map; a naming convention attaches a label to each element of the planar subdivision. Whenever one is afforded the luxury of dealing with a sorted data structure, one may invoke the power of binary search to achieve $O[\log n]$ time complexity during processing of a topographical query, where n is the number of contours which comprise a specific element of the planar subdivision. A topographical query is a request by a user to interpret the position of an arbitrary map coordinate, called the query point, in the context of a topographic map background. An "interpretation" as currently defined consists of a five-tuple of information: the label of the map subdivision element within which the query point resides; the topographical contour of the subdivision element within which the point minimally resides; the local slope at the point; the local elevation at the point; and the flank of the partition element (hillside) upon which the point is situated. As an example, the following list is an interpretation of a query point: (Mt. Hood subdivision, contour #10600-d, 65 degree gradient, 10655 feet elevation, 350 deg NW). As is the case with one-dimensional binary search, the two-dimensional version must concern itself with items having identical keys. Because topographical maps may contain multiple contours lying at the same elevation, the topographical query process must have a mechanism for choosing among them before proceeding. Thus, when "halving the interval", one must check to ensure that the interval is in fact uniquely defined. Inclusion testing within contours is achieved with a deterministic point-in-polygon algorithm developed by the Army in previous research. The traditional normal vector is utilized extensively by the point-in-polygon algorithm, and also by the processing components which interpolate slope and elevation, and determine hillside emplacement. A new theorem derived from the law of cosines provides a decision rule based on integer arithmetic to decide which segments of a polygonal boundary justify a computation of the normal vector. As a byproduct of the research, an algorithm based on the Cevian formula has been developed to find the nearest segment to a query point, without using any floating point operations whatsoever. Future research issues include the topic of visual line-of-sight (binary map coloring) from a query point, and an analysis of the space and time complexity of the new technique, vs. the more burdensome alternative of storing elevation and slope data in large raster archives.

I. BACKGROUND AND TERMINOLOGY

Topographic Line Maps.

A topographic line map (TLM), also known as a contour map, is a vector representation of the boundaries of cross sections of the earth's surface. The projection of the boundary of a cross section of the earth's surface onto the xy -plane is called a topographic *contour*. The equispacing between cross sections along the z axis is called the *contour interval* of the map. A contour map is bounded by the rectangular edges of a map sheet. The edges of a map

form what is called a *clipping region*, which prevents an observer from knowing the behavior of portions of contours which pass outside the rectangular perimeter of the map.

Heuristics to guide map understanding soon become apparent to a novice: e.g., when contours are closely packed together in the xy-plane, the underlying terrain is steep; when far apart, the terrain is flat, or of gentle gradient. As a general rule, contours do not cross, although there are rare exceptions such as natural bridges, overhangs, or bizarre sandstone formations like those found in Utah. For those well-versed in their interpretation, topographic maps can provide a realistic portrayal of actual terrain. However, many human beings find topographic maps difficult to interpret, and have problems visualizing terrain from contour data. It is for this reason that the Army Topographic Engineering Center has opted to utilize perspective displays as an alternative to topographic maps [T1]. Perspective displays render an artistic version of a terrain as it appears from some vantage point near or upon the ground. Figure 1 is a graphic borrowed from reference [K4], and portrays a perspective display of a terrain, together with its corresponding topographic contour representation.

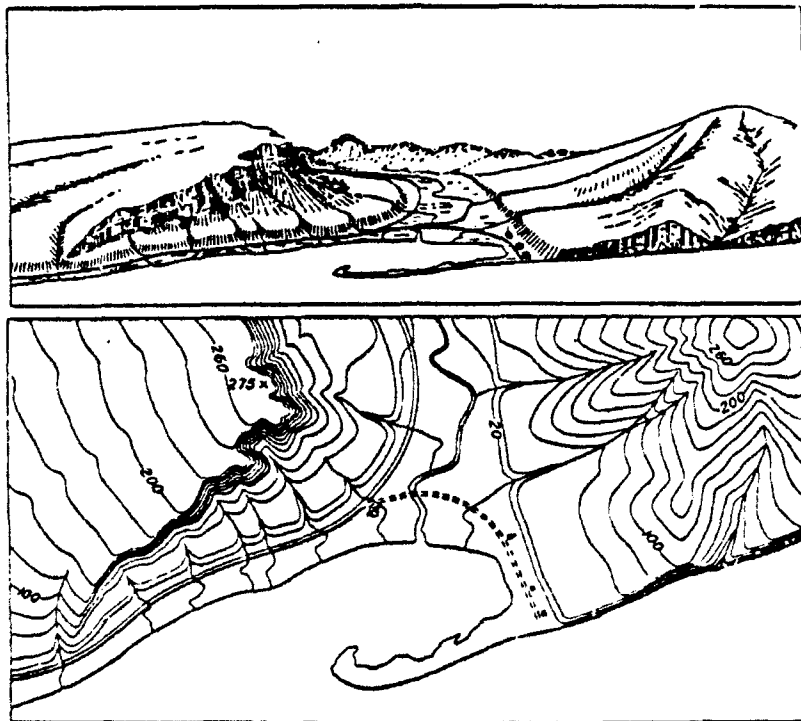


Figure 1. A perspective display and its corresponding topographic line map.

Interpreting a Topographic Line Map.

The objective of automated topographic map interpretation is to write a computer algorithm which locally describes a random query point in the context of a contour map. An interpretation as currently defined consists of five pieces of information: the label (if it exists) of the subdivision within which the query point resides; the name of the topographic contour which brackets the query point from outside; the elevation above or below sea level at the query point; the slope of the terrain at the query point; and the directional gradient at the query point. As an example, the following list constitutes an interpretation of a query point: (Mt. Hood subdivision, contour #10600-d, 65 degree gradient, 10655 feet elevation, 350 deg NW).

The five pieces of information currently sought by the algorithm hardly constitute a complete interpretation of a query point. Other descriptors are desirable in the long term: e.g., the visual line-of-sight from the query point, the profile of a traversible path passing through the query point, the profile of a path which optimally avoids the query point, the feasibility of using the query point as a site for sensor placement, etc. However, the five primitive data currently being returned by the algorithm go far toward providing inputs to some of the higher level queries, which may be synthetically constructed from the primitive queries.

Contour Notation.

In this section, notation is adopted to facilitate reasoning with topographic contours. Associated with every contour is a specific value, denoted $El(C)$, which represents the contour's elevation above or below sea level. The elevation is modulo k , where k is the fixed contour interval of the topographic map. On a particular map, there may be several distinct contours with the same elevation value; for spatial reasoning applications it is important to differentiate among them. Each contour with an elevation value above sea level is contained within another contour, and may itself contain contours.

If contour C_1 is contained within contour C_2 , then C_1 is said to be *nested* within C_2 . A contour cannot be contained within two or more contours which are not nested, but it *can* contain multiple contours which are not nested. If C is a contour of interest, then we denote the contour which minimally contains C to be C^- . By minimally contained, it is meant that any other contour D other than C^- which contains C also contains C^- , which implies that both C^- and C are nested within D . A contour minimally contained by C is called C^+ , where the set of all such contours is denoted $\{C^+\}$. If a query point lies between two elements of $\{C^+\}$, then it is said to be a *saddle point*. Note that $\{C^+\}$ may be the null set. A graphic illustrating these concepts is shown below.

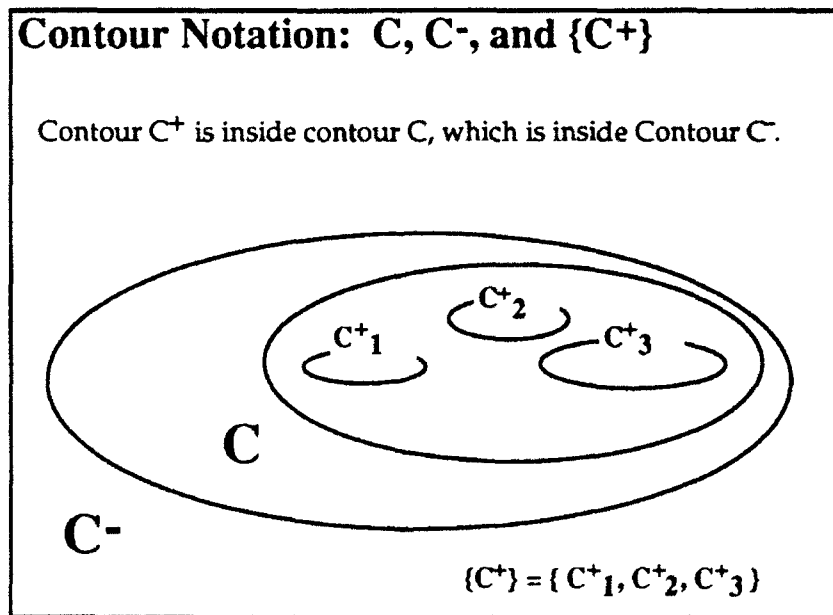


Figure 2. An illustration of contour containment.

A query point and its bracketing contours.

A query point is defined to be any random point of interest. Except for special cases, a query point is bracketed by adjacent contours of a map: one contour which encloses it, called the *outer bracket*, and another contour which does not enclose it, called the *inner bracket*. Since contours are well-ordered at equispaced elevations, the difference in elevation between bracketing contours is equal to plus or minus the fixed contour interval of the map (except for a zero difference at saddle or culvert points). The figure below illustrates the brackets of a query point.

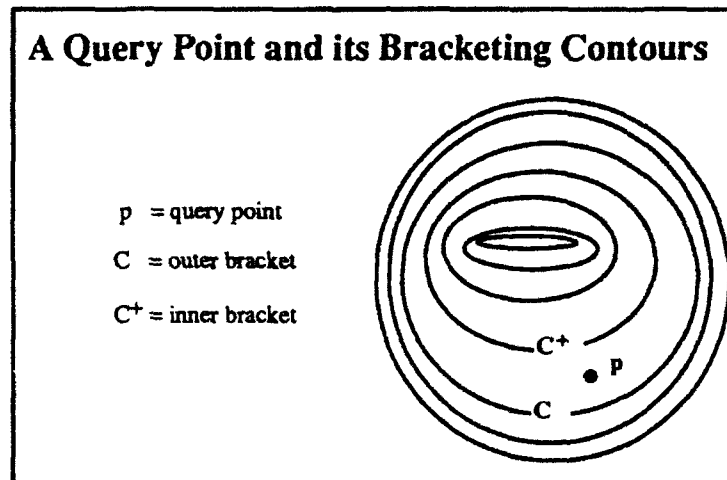


Figure 3. Bracketing a query point from within and without.

II. TWO-DIMENSIONAL BINARY SEARCH APPLIED TO TOPOGRAPHIC MAPS.

Extending binary search to two dimensions to interpret topographic maps.

Binary search has traditionally been applied to a one-dimensional data structure, sorted by some user-defined ordering property. The data structure might be an array of numbers sorted by the natural ordering of the reals, or a list of employee records sorted alphabetically by name. One commonly utilized data structure is 2D trees, in which the data consists of a set of ordered pairs of integers. In a 2D tree, the data is sorted on two keys (the abscissa and the ordinate), with one key primary. A 2D tree is not a true instance of two dimensional binary search data structure, because one key is predominant over another during the sorting process. A better candidate is outlined at [K2], in which an interior point method for linear programming "halves" an ellipse during point-in-polygon testing. However, to be truly elegant, two dimensional binary search should avail itself of the natural containment property inherent to two dimensions. In the digital domain of the computer, two dimensional objects are in general polygons. Just as the one dimensional version must check to see if a point lies between two other points, the two-dimensional version is required to decide if a polygon is contained "between" two other polygons [C1]. Betweenness is equivalent to bracketing a query point with nested polygons.

Topographical contours exhibit a natural ordering due to the way in which the forces of nature have combined to stabilize the crust of the earth. For example, gravity has assured that the top portion of a mountain has

a smaller cross section than its base. Thus, when projected onto a plane, contours from the same mountain appear to be nested. Ordering by elevation, and nesting by containment are properties which may be exploited to sort contours. The data structure which results by appealing to a two-dimensional sort on elevation and nesting is called the *contour containment graph*. The motivation is that to exploit the $O[\log n]$ query power of binary search, one requires that the underlying data structure be sorted. We will see below that there are two preprocessing steps required to set up an efficient two-dimensional search of topographic maps: the first is the construction of the contour containment graph, and the second is the partitioning of the containment graph into regions suitably indexed for binary search.

The Contour Containment Graph, and Labeling of Topographic Features.

As a first step in constructing the contour containment graph, we can uniquely label each contour, and then sort all contours on elevation, in ascending order. We then "nest" contours. To illustrate, suppose a specific 100-meter contour is labeled, and the contour interval of the map is 10 meters: we now seek to find all 110 meter contours contained within the labeled contour. If we find one, we create a pointer from the 100-meter contour's label to the label of the 110 meter contour discovered to be contained in the contour. We continue this process until no more contours are found to be within the 100-meter contour. We repeat this operation for all other 100-meter contours. When this step is completed, we switch our baseline cell complex from all those bounded by 100-meter contours to all those bounded by 110-meter contours, and continue the process until there are no contours remaining to be processed. An example of a terrain and its contour containment graph is depicted in the figure below. The terrain features three hills. All three are contained within baseline contours of twenty and forty meters elevation. Note that a label may be associated with the forty meter contour to delimit the extent of the "hill country". Also, a label may be installed on each of the sixty meter contours to name the individual hills. One of the three hills contains two small knobs at the top, at an elevation of one hundred twenty meters. Each time that the set $\{C^+\}$ contains multiple elements for a given contour C , another level of sorting must be initiated to assure that the contour containment graph is properly structured and nested for binary search.

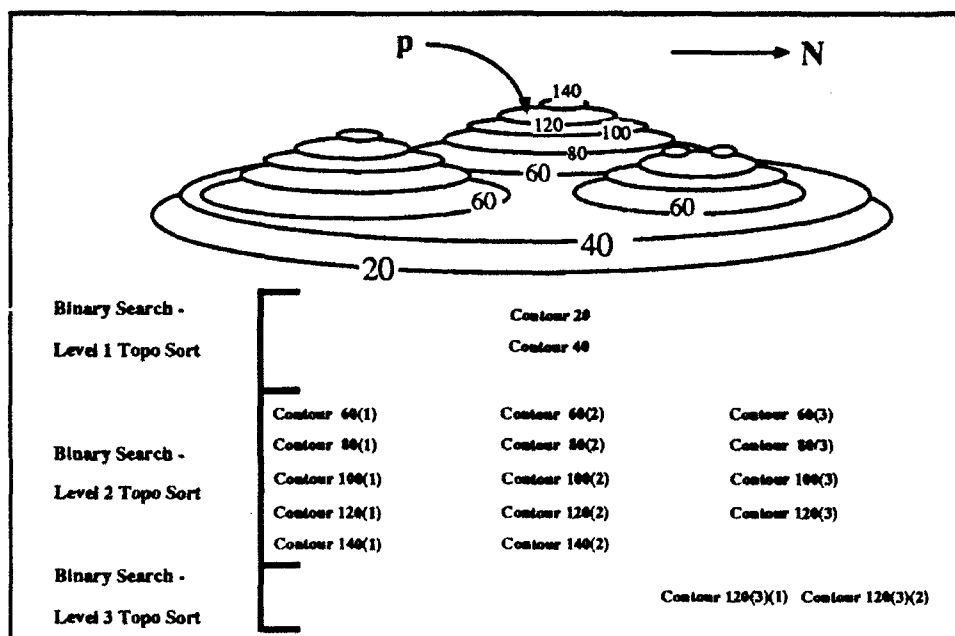


Figure 4. A sorted terrain, nested in preparation for binary search.

Within each level of a contour containment graph, binary search may be invoked to achieve $O[\log n]$ time complexity, where n is the number of contours contained at that level. To illustrate, in the figure below, a hill is represented by eight contours. On the first iteration of binary search, a contour halfway up the hillside at eighty meters is considered, and the query point is determined to be inside. On the second iteration, it is determined that the query point is not inside the one hundred twenty meter contour, which is three quarters of the way up. The third iteration decides that the point is not inside the one hundred meter contour, which is five eighths of the way to the top. At the next iteration binary search concludes, having bracketed the query point between the eighty and one hundred meter contours, while having interrogated only three of the eight contours.

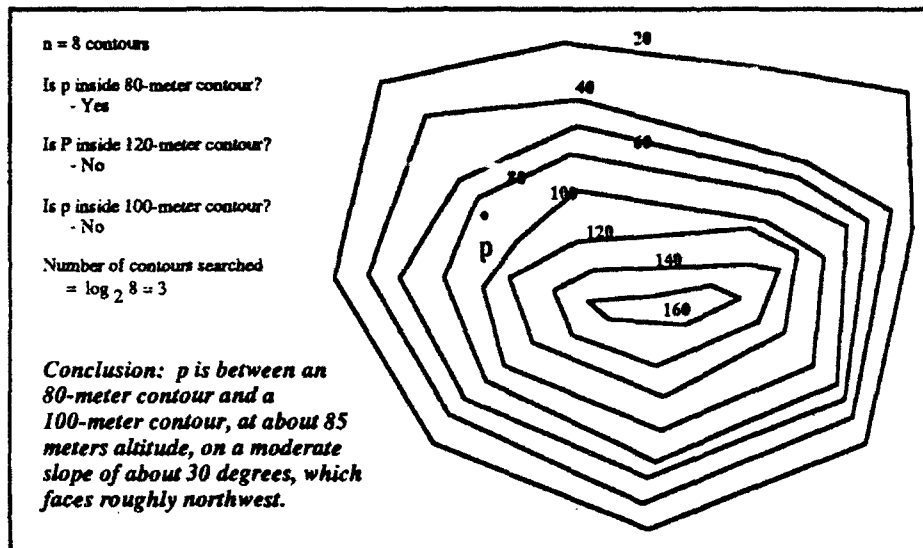


Figure 5. Binary search brackets a query point.

Although two-dimensional binary search may achieve $O[\log n]$ time complexity over a database of n contours, the issue remains open regarding the time complexity of the search as a function of the number of vertices contained within a given contour. For example, one topographic contour may contain a single vertex, whereas another may contain thousands of vertices. Processing a set of contours comprised of a small number of vertices is clearly more desirable for performance considerations than processing a set of contours comprised of a large number of vertices. An objective metric of time complexity should take into account both the number of contours and the number of vertices per contour.

Partitioning a Topographic Map for Binary Search.

Any topographical map contains contours of locally minimum elevation. These are readily identified from the contour containment graph developed in the preceding section. The strategy is to partition the map between all such contours, by constructing synthesized boundaries to act as cuts for binary search. Optimal placement of the cut boundaries is a load balancing problem, which needs to address not only the number of contours within each cut, but the total number of vertices which comprise contours in the cut. In the diagram below, four hills have been partitioned by synthesized boundaries into regions suitable for binary search. Note that the bold lines are not contours but synthetic boundaries. The first cut runs roughly down the middle of the map, and segregates the rightmost hill from the other three. Observe that the first cut contains nine contours on the left, but only seven on the right. This is not arbitrary, but is designed to compensate for the longer perimeters of the contours on the right of the cut (it is implied that a longer perimeter equates to a larger number of vertices in the contour boundary data).

The second cut is dependent upon the decision made during the first cut. If a query point is to the left of the first cut, then the second cut lies between the two most northerly hills and the hill in the southwest corner. Conversely, if the query point is to the right of the first cut, then the second cut lies halfway up the rightmost hill. Continuing in this fashion, the number three cuts are synthesized. No further cuts are shown, but the logic to create them is similar.

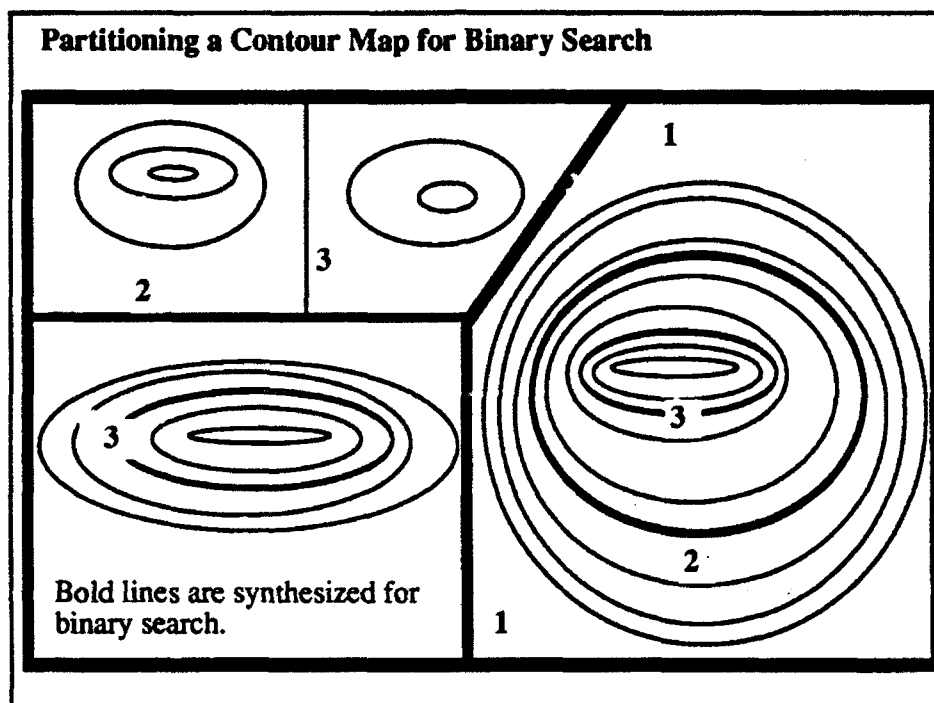


Figure 6. Load balancing a contour map to create a two-dimensional binary tree.

Dealing with contours which exit the clipping region of a map.

Figure 6 is oversimplified. In general, contours are not so well-behaved. There is one common problem to consider: a contour may exit the rectangular region bounding the map, and therefore pass outside the clipping region. The problem may be solved by conjoining the troublesome contour with the rectangular edge of the map. This contrivance forces two polygons to be synthesized from the errant contour, to create a data structure compliant with two-dimensional binary search. Synthetic boundaries for binary search may also be constructed accordingly.

The figure below depicts a clipped contour of forty meter elevation which exits the map at both sides. Two dimensional binary search requires that data structures be in the form of polygons. We synthetically create two new polygons by conjoining the clipped contour with the edges of the map. Because the point-in-polygon algorithm of choice (described in the next section) requires a sense of handedness, we assure that the vertices of the new polygons are in counterclockwise order. At execution time, we may now ask if a query point is contained within either the upper or the lower polygon manufactured by utilizing the clipped contour, and proceed accordingly.

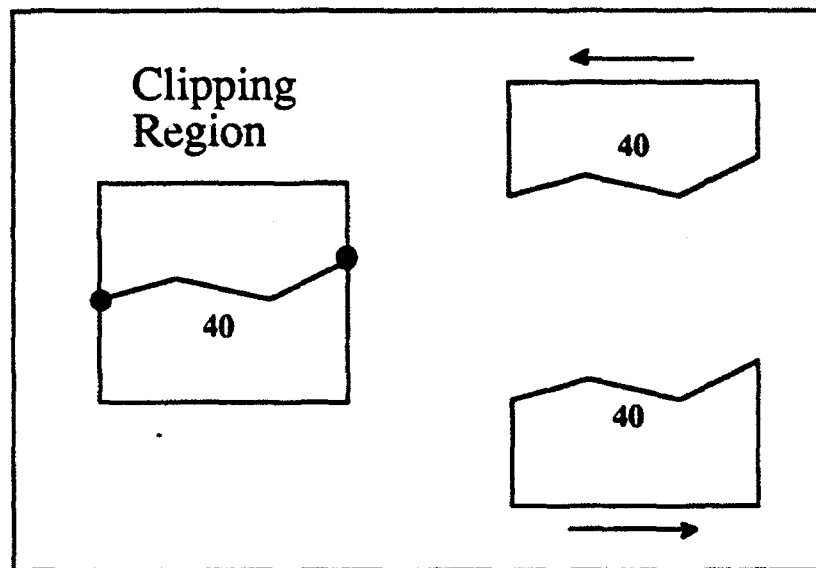


Figure 7. Creating two polygons from a clipped contour.

III. AN INCLUSION (POINT-IN-POLYGON) ALGORITHM, AND PROXIMITY.

Perceived shortcomings of currently available point-in-polygon technology.

The two-dimensional binary search algorithm requires a utility function to establish whether or not a query point is inside a topographical contour. The utility function is a true workhorse, so it must be efficient. There is no margin for error, which means that point-in-polygon algorithms which rely on the precision of machine arithmetic are inappropriate candidates. For this reason, approaches based on the winding number or the parity algorithm are currently infeasible. The Apple Macintosh family of computers has implemented a predicate called "point-in-region-p", available as part of the Quickdraw graphics repertoire, but the predicate consumes quadratic amounts of region space in memory, which becomes prohibitive for even a moderate number of polygonal boundaries. A high-performance algorithm from the computational geometry literature, based on triangulation [K3], is a viable candidate, although it remains an untested quantity, since it has never been tasked against a multi-megabyte database of topographical contours.

Because of perceived shortcomings of on-the-shelf point-in-polygon algorithms, and the lack of benchmark data to test the performance of the triangulation algorithm, the author has opted to implement his own algorithm [C2], which has been extensively tested against actual contour data. The algorithm assures that a contour is oriented in a counterclockwise direction, so that the interior of the contour is to the left during traversal. Inclusion testing is then conducted as a function of a query point's proximity to a contour (see figure below). One benefit of the algorithm is that it returns distance and direction (normal vector) to the nearest point on a boundary, in addition to the inclusion decision. As will be seen below, the normal vector is crucial to topographic map interpretation. As originally conceptualized, the algorithm anticipated that every pixel in a digital boundary would be explicitly available as part of the data structure. However, the Defense Mapping Agency does not represent feature boundaries so obviously. Instead, a contour is provided in chain-coded format, where the boundary of the contour consists of a set of ordered vertices. It is up to the user of the data to create the edges which join the vertices.

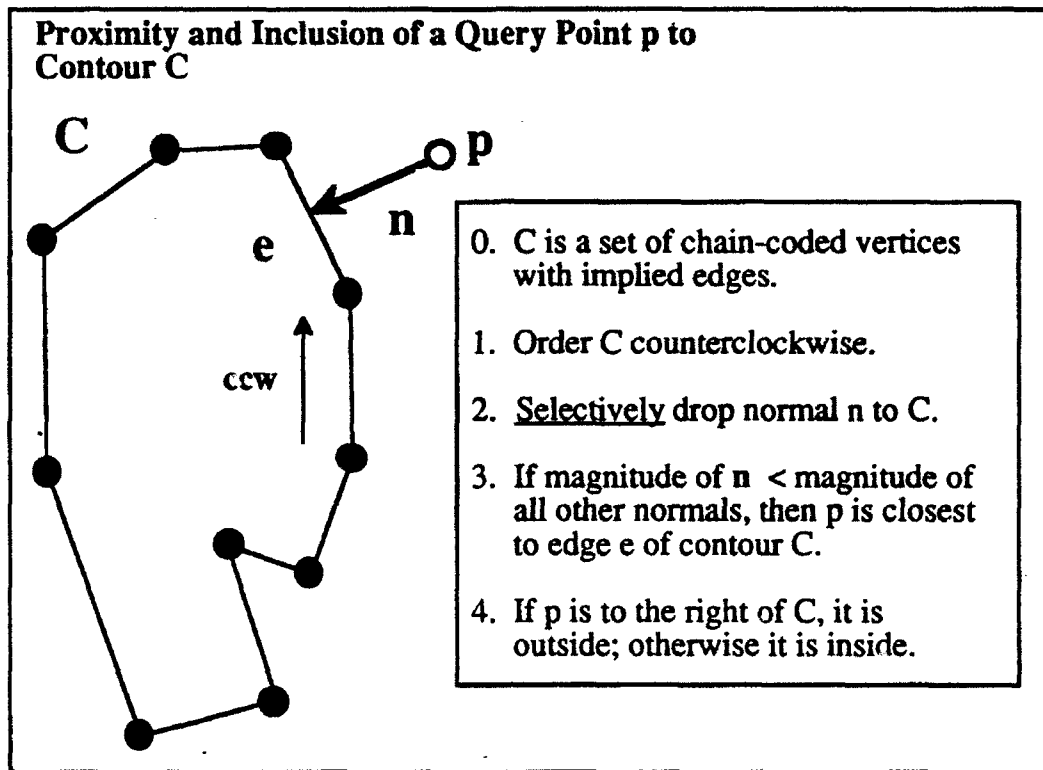


Figure 8. The normal vector may be used to decide inclusion.

The Voronoi diagram for data produced by the Defense Mapping Agency.

Vector data distributed by the Defense Mapping Agency (DMA) contains three kinds of objects: points, line segments, and polygons. It has been known for some time that the skeleton, or medial axis of a polygon consists of portions of parabolas and line segments [B2]. The parabolas are the locus of equal distance between points and segments. The line segments are the locus (angle bisectors) between extended segments. It is also true that for any set of points, segments, and polygons the equidistance locus consists of parabolas and line segments. Thus, the Voronoi diagram for DMA data, which is defined to be the locus of equal distance, is in general parabolic.

There is currently no commercial product available to generate the parabolic Voronoi diagram for an arbitrary set of polygons, segments, and points. However, there are three research and development tools (of varying degrees of robustness) circulating among researchers in academia [M3]. The developmental products implemented to date have encountered problems of numerical precision, primarily when deciding upon which side of a parabola a query point lies [F2]. However, as indicated at reference [E1], the theory behind the sweepline algorithm [F1] to generate the linear Voronoi diagram should be directly extensible to the parabolic diagram. It is clear that for the asymptotic solution to the static proximity problem, the Voronoi diagram is the paradigm of choice. As a stopgap measure, until a tool to generate the parabolic diagram is available, the author has developed his own proximity algorithm, described below, based on restricted use of the normal vector. The author's algorithm, unlike the Voronoi diagram, facilitates dynamic objects. If an object's position changes, the Voronoi diagram must reinvolve a relatively expensive preprocessing step, whereas the author's algorithm simply replaces the object's old boundary position with the new.

Finding the nearest point of a contour to a query point.

A contour, which when represented with digital data is in the form of a polygon, consists of a set of vertices and the implied edges which connect the vertices. Thus, when one speaks of proximity to a contour from a query point, one is actually referring to minimal Euclidean distance to the set of vertices, vs. distance to the set of edges.

Minimal distance to an edge is non-trivial to compute. This process entails dropping the normal vector from a query point to the edge. Since floating point operations may be required at every edge to which the normal is dropped (although the author introduces below a new technique which avoids floating point arithmetic), we would like to limit the number of edges incurring such an expensive operation. If the normal vector strikes an edge directly, the edge is said to be *admissible* to the normal vector. Refer to the figure below. Clearly, it does not behoove us to drop the normal from query point p to edge e_2 , since the tip of the normal does not even intersect e_2 , but rather its extension. Such cases are precisely those which we strive to avoid, by appealing to a normal vector admissibility filtering technique. It will be shown below that as a side effect, the filtering technique returns minimal distance to a vertex.

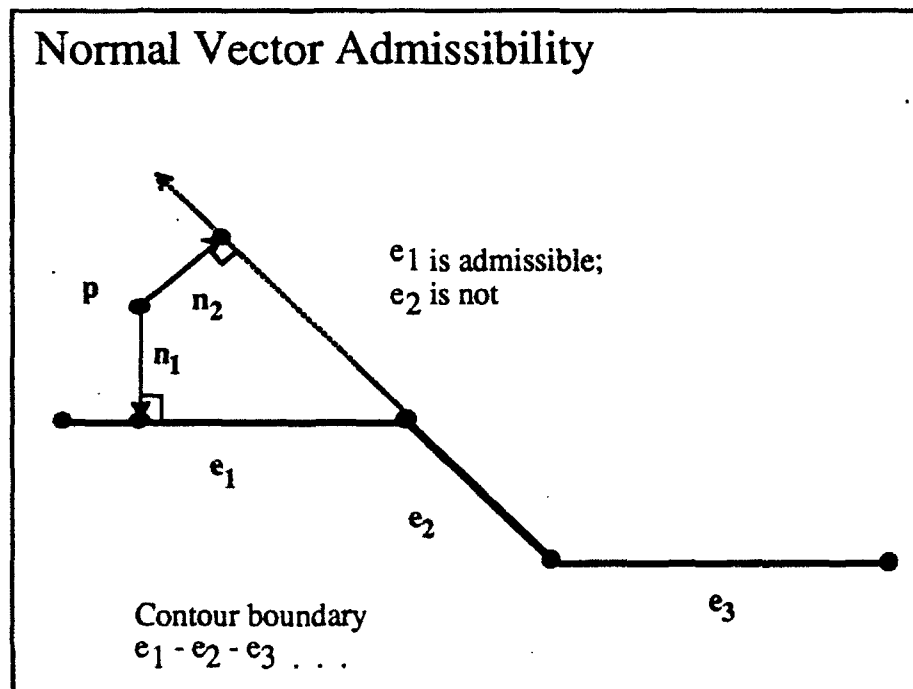


Figure 9. Certain contour edges do not admit the normal vector.

Derivation of edge admissibility conditions from the law of cosines.

Construct orthogonal rays from the endpoints of contour edge e , as in Figure 10 below. Now suppose that query point p lies between the rays. Note that the angle between edges x and e is acute, as is the angle between edges y and e . Let the angle between y and e be θ_1 and the angle between x and e be θ_2 . Then, by the law of cosines,

$$\begin{aligned} x^2 &= y^2 + e^2 - 2 y e \cos \theta_1 & [1] \\ y^2 &= x^2 + e^2 - 2 x e \cos \theta_2 & [2] \end{aligned}$$

The cosine function is positive for acute angles. We therefore obtain

$$x^2 + \alpha_1 = y^2 + e^2 \quad [3]$$

$$y^2 + \alpha_2 = x^2 + e^2; \quad \alpha_1, \alpha_2 \geq 0 \quad [4]$$

These equations are alternatively expressed by the inequalities:

$$x^2 \leq y^2 + e^2 \quad [5]$$

$$y^2 \leq x^2 + e^2 \quad [6]$$

This set of inequalities must be true for segment e to admit the normal vector. Point p of Figure 10 satisfies the conditions.

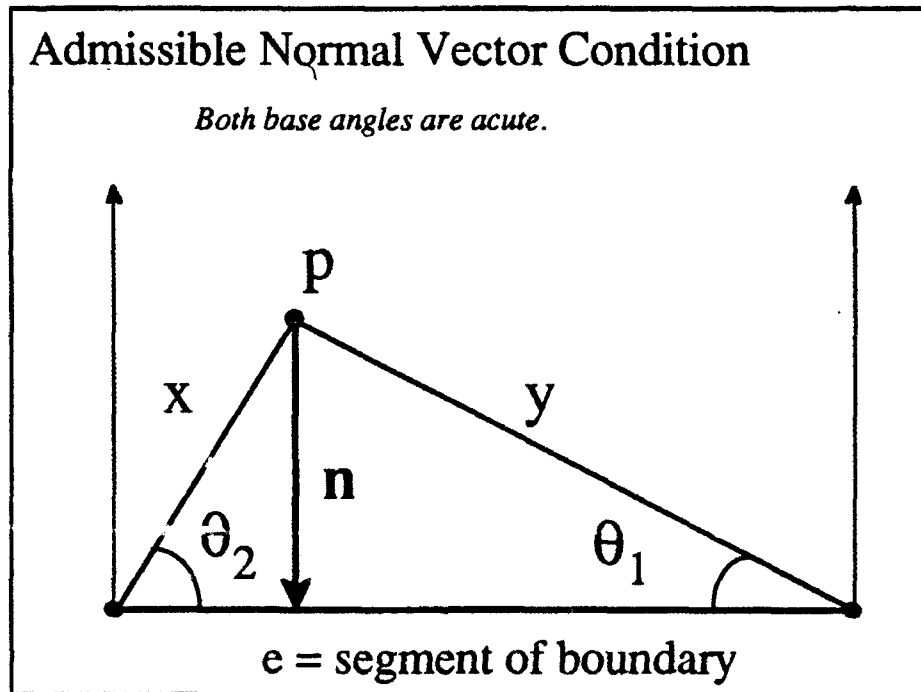


Figure 10. An edge is admissible if base angles are both acute.

In practice, it is more likely for the test to be failed than to be passed, so it makes sense to test first for failure rather than for success. The failure condition may be written as the predicate

$$\neg [x^2 \leq y^2 + e^2 \quad \wedge \quad y^2 \leq x^2 + e^2] \quad [7]$$

From DeMorgan's rules, this may be rewritten

$$\neg [x^2 \leq y^2 + e^2] \quad \vee \quad \neg [y^2 \leq x^2 + e^2] \quad [8]$$

which is equivalent to

$$x^2 > y^2 + e^2 \quad \vee \quad y^2 > x^2 + e^2 \quad [9]$$

If either side of disjunction [9] is true, then edge e is not admissible to the normal vector, and a potentially expensive floating point operation is avoided by means of a simple integer-valued decision function. An example of satisfaction of the second inequality of the disjunction is illustrated at the figure below. In this case, edge e fails the admissibility condition, so that the normal vector computation is avoided.

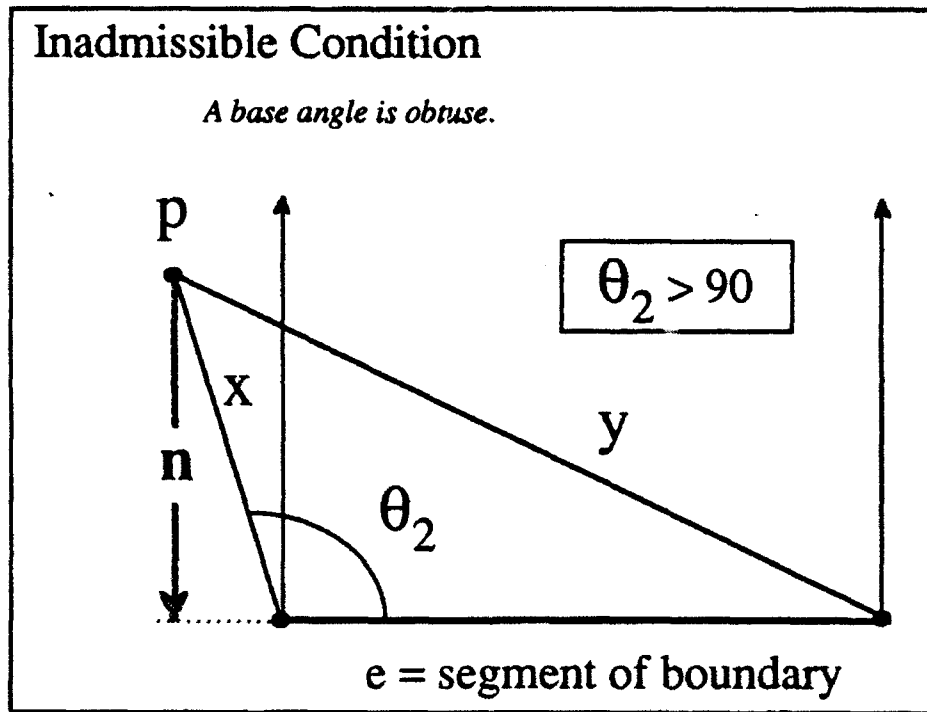


Figure 11. An obtuse base angle precludes admissibility.

As a byproduct of the admissibility test, minimal distance to a vertex is returned. Consider the integer-valued expression $(s_p - s_v)^2 + (t_p - t_v)^2$, where (s_v, t_v) is the coordinate at the vertex and (s_p, t_p) is the coordinate at the query point. This expression is synonymous with either of the arguments x^2 or y^2 in equations [1]-[9] above. Hence, the filtering operation as a side effect monitors the squares of the distances to each of the vertices of a contour. When the smallest such expression is found across all vertex possibilities, the square root is extracted. The entire process involves n integer-valued operations for n vertices, and one floating point operation, for a time complexity of $O[n]$. The integer-valued operation here involves two integer multiplies, three integer adds, and an integer comparison. The floating point operation is a single-shot appeal to the square root of the minimal integer-valued result.

A Common Lisp implementation of the edge admissibility test might appear as follows:

```
(defun admissible-normal-segment-p (x y ax ay bx by)
; (ax, ay) and (bx, by) are the endpoints of segment e in figures; (x,y) is query point.
  (prog (dis1sqr dis2sqr dis3sqr)
    (declare (type longint x y ax ay bx by dis1sqr dis2sqr dis3sqr))
    dis1sqr = (dissqr ax ay bx by)
    dis2sqr = (dissqr x y ax ay)
    dis3sqr = (dissqr x y bx by)
    (cond ((> dis3sqr (+ dis1sqr dis2sqr))(return nil))
          ((> dis2sqr (+ dis1sqr dis3sqr))(return nil))
          (t (return t)))))
```

Finding the normal vector with minimum magnitude, across all segments.

Although we now have a test to determine which segments of a boundary admit the normal vector from a query point, we have not said anything about the actual computation of the minimal such vector across all segments. In this section we develop a new test to find the smallest normal vector, without resorting to any floating point computations. If the actual magnitude is desired, two floating point operations are required over the entire database. We appeal to a very useful result from analytic geometry, called the Cevian formula (for a development see [K1]). A *cevia* is defined to be a line segment drawn from a vertex of a triangle to the opposite side. Note that medians, angle bisectors, and altitudes are all examples of cevians. The Cevian formula is shown in the figure below, where n is an altitude in this case. It is convenient that the altitude is equivalent to the normal vector under discussion. In the figure, observe that r_z and s_z are lengths which sum to side z , whereas r and s are ratios which sum to one.

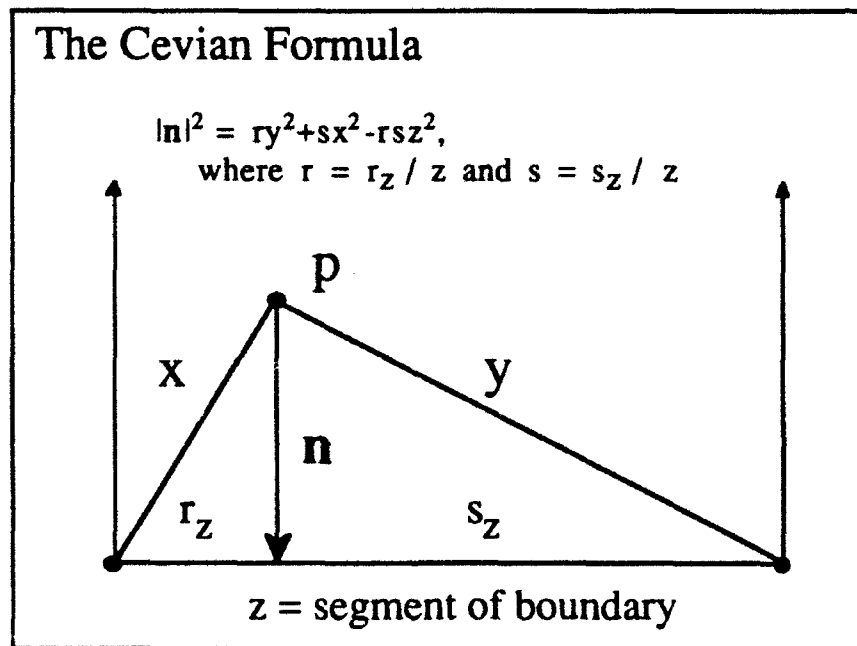


Figure 12. The Cevian formula relates a normal vector to the sides of a triangle.

Unfortunately, we do not know the values of r and s , because we do not know the point at which the normal vector impacts side z . In the equations below, which until step [15] echo the discussion in [K1], we derive a formula for the square of the magnitude of the normal in terms of a ratio involving the squares of the sides. Steps [10]-[11] are a reiteration of the information conveyed by the figure. Steps [12]-[13] involve a substitution for s , followed by a reformulation as a quadratic equation in terms of r . In step [14] we set the discriminant equal to zero, because the roots of equation [13] must be non-negative and equal, since r and s form a convex set. Solving this equation for n^2 results in the ratio shown at [15], which but for the divide operation is economical to compute, since it involves four integer multiplies and three adds. If one were tuning the technique with assembly code, two of the multiplies (those involving the 4) could be converted into two-bit left shifts, since shifts are cheaper than multiplies.

$$r = \frac{r_z}{z}; s = \frac{s_z}{z}; r + s = 1; r_z + s_z = z \quad [10]$$

$$n^2 = ry^2 + sx^2 - rsz^2 \quad [11]$$

$$n^2 = ry^2 + (1-r)x^2 - r(1-r)z^2 \quad [12]$$

$$z^2r^2 - (x^2 + z^2 - y^2)r + (x^2 - n^2) = 0 \quad [13]$$

$$(z^2 - y^2 + x^2)^2 - 4z^2(x^2 - n^2) = 0 \quad [14]$$

$$n^2 = \frac{4x^2z^2 - (x^2 + z^2 - y^2)^2}{4z^2} \quad [15]$$

What about the division by $4z^2$, which implies a floating point operation? The answer is that in order to find the normal vector of smallest magnitude, we may refrain from performing the division until all admissible segments have been associated with a numerator and denominator as at [15], and checked against the shortest normal vector found thus far. The check is made as follows. Let n_1 be the normal dropped from a query point to segment z_1 , with x_1 and y_1 the distances to the respective endpoints of z_1 . Let n_2, z_2, x_2 , and y_2 be defined similarly. Then the squares of the normals are shown respectively at [16] and [17]. Now $n_1 < n_2$ if and only if [19] and [20] are true, but [20] is true if and only if the product of the means is less than the product of the extremes as shown at [21]. Cancelling the common factor produces test inequality [22]. Notice that if we are using integer-valued coordinates, as we must if we are working with data displayed to a computer screen, there are no floating point expressions involved in the test.

$$n_1^2 = \frac{4x_1^2z_1^2 - (x_1^2 + z_1^2 - y_1^2)^2}{4z_1^2} \quad [16]$$

$$n_2^2 = \frac{4x_2^2z_2^2 - (x_2^2 + z_2^2 - y_2^2)^2}{4z_2^2} \quad [17]$$

$$n_1 < n_2 \Leftrightarrow \quad [18]$$

$$n_1^2 < n_2^2 \Leftrightarrow \quad [19]$$

$$\frac{4x_1^2z_1^2 - (x_1^2 + z_1^2 - y_1^2)^2}{4z_1^2} < \frac{4x_2^2z_2^2 - (x_2^2 + z_2^2 - y_2^2)^2}{4z_2^2} \Leftrightarrow \quad [20]$$

$$[4x_1^2z_1^2 - (x_1^2 + z_1^2 - y_1^2)^2]4z_2^2 < [4x_2^2z_2^2 - (x_2^2 + z_2^2 - y_2^2)^2]4z_1^2 \Leftrightarrow \quad [21]$$

$$z_2^2[4x_1^2z_1^2 - (x_1^2 + z_1^2 - y_1^2)^2] < z_1^2[4x_2^2z_2^2 - (x_2^2 + z_2^2 - y_2^2)^2] \quad [22]$$

Using the test is simple. As input we receive a query point and candidate segment with integer coordinates. The squares of the distances from the query point to the segment endpoints are computed with the usual Euclidean formula, as is the square of the distance between the endpoints. These three quantities are used to compute the integer-valued numerator and denominator of equation [15]. The same technique applied to some other candidate segment produces another numerator and denominator, which we cross-multiply with the first at inequality [22]. If the product of the means is less than that of the extremes, then the first segment is closer to the query point; otherwise the second segment is closer. We continue this process until all segments are exhausted, remembering the segment giving rise to the shortest normal vector as we do so.

Observe that we have located the nearest segment (according to the true Euclidean metric) to a query point without resorting to any floating point arithmetic. Granted, we do not yet know the magnitude of the shortest normal vector, but we know that we have the shortest. To obtain the magnitude, we merely need to perform the division indicated at equation [15], and extract the square root of the result. Note also that we never had to compute any of the points of a line segment; we were able to make do with the vertices at its endpoints. This latter artifact demonstrates the power and leverage of the Cevian formula, developed over three centuries ago. The formula may potentially be used to assist in the generation of the parabolic Voronoi diagram for line segments and polygons.

We briefly summarize before moving on to the next section. When the two-dimensional binary search paradigm requests the inclusion algorithm to decide whether or not a query point is contained within a specific contour, the inclusion algorithm is handed the counterclockwise-oriented set of contour vertices and the query point as arguments. The first action taken by the inclusion algorithm is to subject all of the implied edges of the contour to the normal vector admissibility test, maintaining the squared distances to the vertices on the side. Generally, the test returns just a handful of edges admissible to the normal vector. To each of these, the cross-product test shown at [22] is performed to locate the minimal normal vector. This quantity is compared against the minimal result obtained for the vertices. If the square of the distance to an edge is smaller than the squared distance to a vertex, a test is invoked to decide if the query point is to the left or the right of the edge; if to the left, the point is inside the contour, and if to the right, the point is outside. At this time the numerator and denominator of equation [15] may be divided and the square root extracted to obtain the actual magnitude of the normal vector. If the squared distance to a vertex is smaller than that to an edge, a synthetic edge is constructed from the vertex's predecessor and successor vertices in the contour boundary, and a test is invoked to decide if the vertex is to the left or the right of the synthetic edge; if to the left, the query point is inside the contour, and if to the right, the point is outside. The square root may be extracted to obtain the magnitude of the normal vector. The shortest normal vector points to either the inner or the outer bracketing contour of the query point.

IV. INTERPRETATION OF A QUERY POINT IN THE CONTEXT OF A MAP.

Binary search of a contour containment graph concludes by returning the two bracketing contours of a query point. The algorithm is now armed with all the information it requires to produce an "interpretation" of a query point, as defined in the first section of the paper. If either bracket has inherited the name of a mountain, hillside, crater, etc., for which the bracket is a structural element, then the name is available for simple display, or for further spatial reasoning operations such as line-of-sight or traversability reasoning. Because inclusion testing as described above returns as a byproduct the normal vector from a query point to a contour, both the distance to the outer bracket and the distance to the inner bracket are known when binary search completes. These two distances may be used in conjunction with the contour interval of the map to obtain estimates for the point's elevation and slope. The direction from a hilltop to the query point, together with the elevation values and orientation of the bracketing contours, may be used to determine a directional gradient, which establishes upon which flank of a hillside a query point resides. The details involved in extracting the elevation, the slope, and the directional gradient are described below.

Deriving the elevation of a query point from its bracketing contours.

Once the bracketing contours for a query point have been established, it is a simple matter to compute an interpolated elevation at the query point. Without loss of generality, let us assume p is on an uphill slope from outer bracket C to inner bracket C^+ , as depicted at the figure below. The elevation of query point p , denoted $El(p)$, may be obtained by using similar triangles to compute an expression which accounts for p 's relative location between the contours, and multiplying it by the fixed contour interval of the map. To this expression is added the baseline elevation at p 's outer bracket (if p were on a downhill slope, the expression would be subtracted instead). Special cases require additional processing. If a query point has an outer bracket but no inner bracket, as it will when it resides within a contour which represents a hilltop, and there are no control points available to indicate the actual elevation at the hilltop, then the query point inherits the elevation of its outer bracket, since interpolation is impossible. If a control point is available (generally obtained by surveyors with a spirit level, and represented on the map with an "X" or a delta symbol), then interpolation is possible even in the absence of an inner bracketing contour. One simply coerces the inner bracketing contour to be the control point, and temporarily sets the map contour interval to be the difference between the elevation of the control point and the elevation of the outer bracket. Downhill slopes, craters, saddles, and culverts may be treated with similar logic.

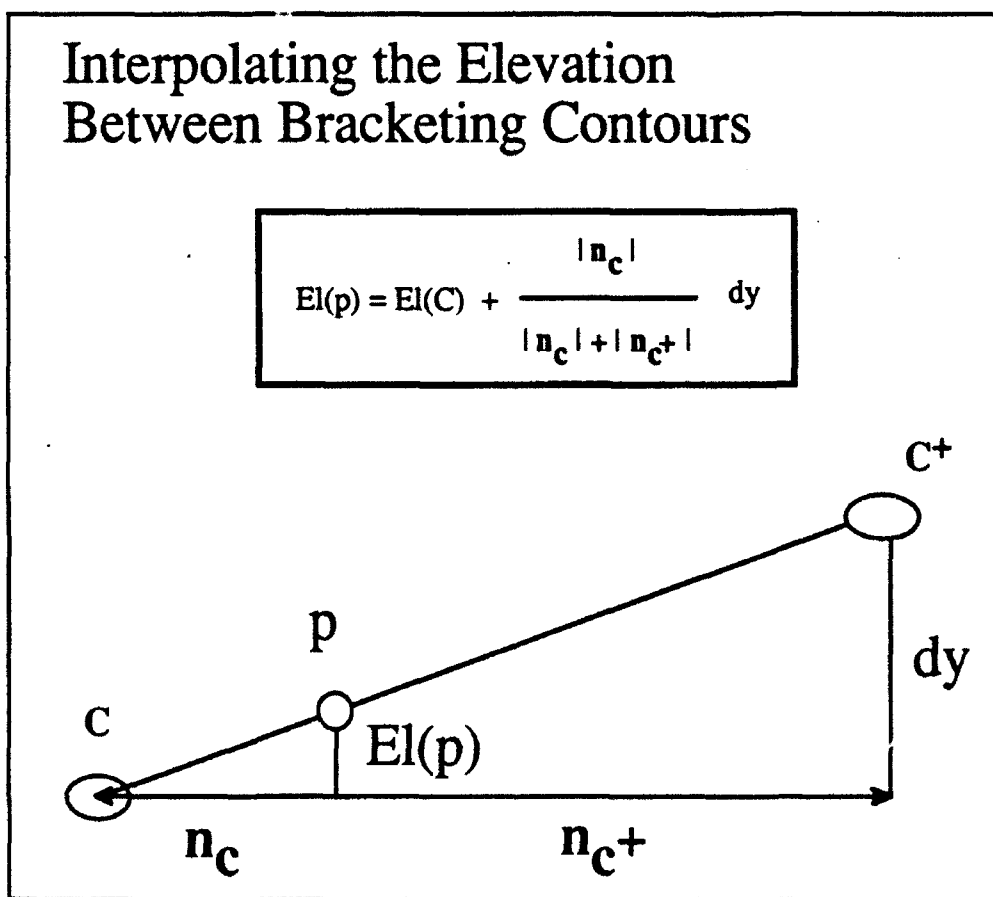


Figure 13. Elevation is obtained through simple linear interpolation.

Obtaining the slope at a query point from its bracketing contours.

The local slope at a query point is so simple to estimate that even interpolation is not required. It is simply the angle with a tangent equal to "the rise over the run". The "rise" is fixed, as it is given by the contour interval of the map. The "run" is defined to be the sum of the magnitudes of the normal vectors drawn to the outer and inner bracketing contours. At the top of a hill or at the base of a depression, in a saddle or a culvert, the slope is assumed to be zero, for flat ground. However, if a control point is available to provide additional elevation data, then logic similar to that outlined for elevation in the paragraph above may be utilized to obtain a refined estimate of slope. Outside the limits of the lowest lying contours, the algorithm is designed to return the string "drainage area", which again is assumed to be flat ground. There may or may not be a perennial stream flowing through a drainage area, but during flashfloods it is assumed that water would flow there.

Note that a peculiar thing happens if we slide the query point along either of the normal vectors pointing to the bracketing contours. The slope remains fixed as we do so. This is the price we pay for approximating a terrain by a set of cross-sectional contours. The computed slope cannot be made more accurate than the resolution imposed by the contour interval of the map. Thus, between any two nested contours, there is a vector field of slope vectors which connect every digital point of the inner bracket with some digital point of the outer bracket, and vice versa.

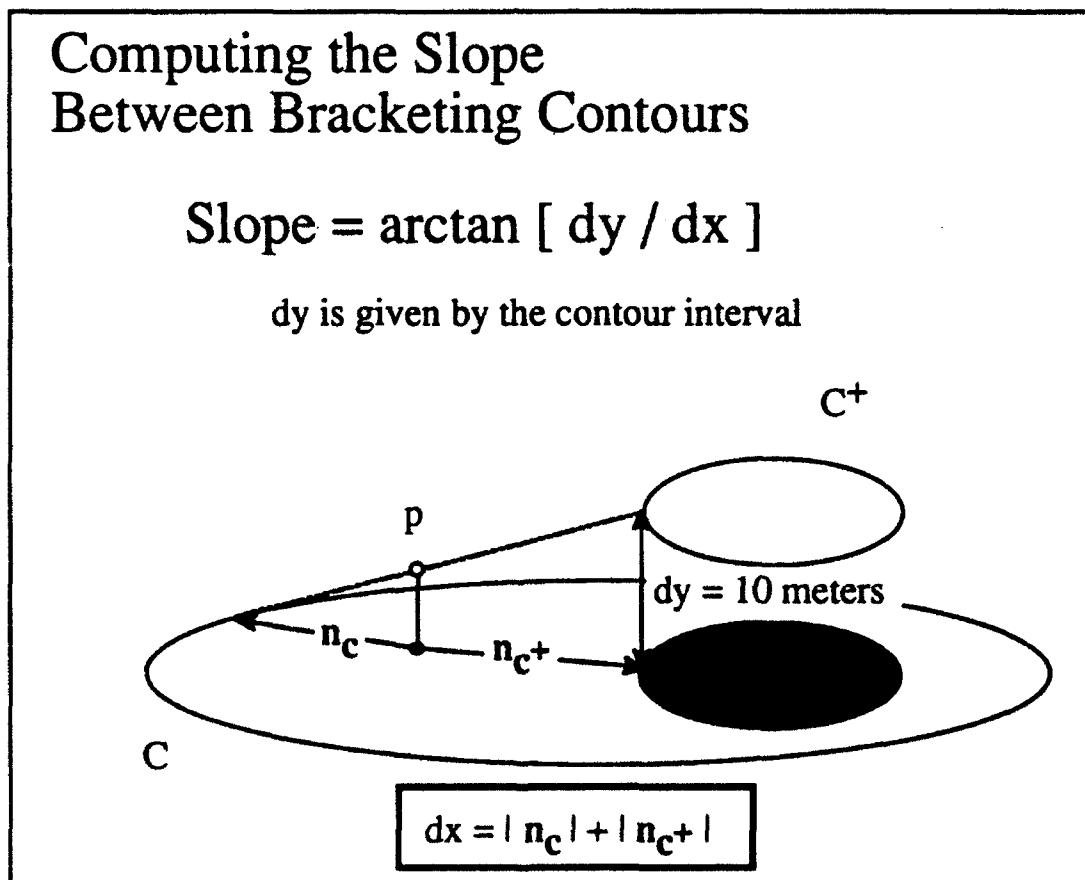


Figure 14. The "rise" is fixed, and the "run" is the sum of the normal vector magnitudes.

Obtaining the directional gradient at a query point from its bracketing contours.

Again, assume the familiar example of an uphill slope, so that $El(C) = El(C^+) - k$, where k is the fixed contour interval of the map. Construct the vector from the hilltop to the query point. Define the hilltop to be the control point at the top of the hill if it exists; otherwise make it some reasonable estimate, such as the centroid of the topmost contour. If there are multiple topmost contours, then make the hilltop the centroid of them all. Suppose that the hilltop to query point vector points to the left, as in the figure below. Then it is pointing downhill, because C 's elevation is less than that of C^+ , and it is pointing to the west, since due north is as shown by the map. The query point is therefore on the western flank of a hillside. Variations on this theme are computable for other configurations of terrain. If the elevation of C is greater than that of C^+ , and we observe a leftward-pointing vector, then we would be on the western flank of a crater or valley. If the elevation of C was to be equal to that of C^+ and the vector was to point to the south, then the query point would be on a saddle or in a culvert, oriented in an east-west fashion.

The vector pointing from a hilltop to a query point is a suitable gauge of directional gradient from a global perspective. However, a query point may be situated locally on a geologic feature of a hillside, with an orientation seemingly at odds with the global result. For example, on the south side of a mountain, there may be a ridge which proceeds from the summit down to the south. The ridge will have both eastern and western flanks. Suppose for the sake of argument that a query point is on the western flank of the ridge. We conclude it is possible for a query point to be locally on a western flank, but globally on the south side of the mountain. The local flank estimate is easily computed by drawing the normal vector from the query point to its outer bracketing contour. The vector points in the compass direction of the local gradient. This procedure is particularly useful for rugged terrain such as that encountered on Mount Rainier in the state of Washington, where contour data tends to resemble a set of nested "octopuses".

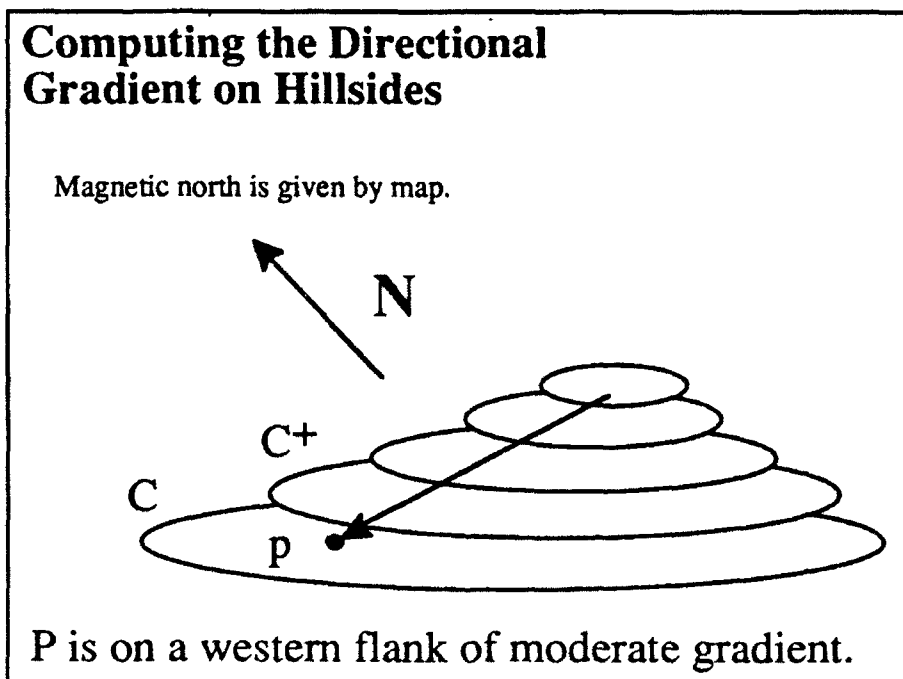


Figure 15. Determining a query point's emplacement on a hillside.

V. AN IMPLEMENTATION, AND CONCLUSIONS.

The theory of automated topographic map interpretation, as developed to date, has been partially implemented on a Macintosh IIx workstation, using Macintosh Common Lisp, version 2.0. There are plans to convert the code into C, using the Symantec Think C environment. The conversion is intended not so much for performance purposes, since the Lisp compiler's performance is favorable when compared to that of the Symantec package, but to be able to control the process of garbage collection, which in the Lisp package is beyond the reach of the user.

There are two databanks of contour data: real and simulated. The first source of the real variety is a set of digital elevation matrix (DEM) data, which is a gridded representation of elevation values sampled at equispaced x and y increments. DEM data is produced by the United States Geological Survey (USGS) office, as a result of data collection performed primarily by civilian engineers. To obtain topographical contours from DEM data, one may utilize a geographic information system (GIS) to extract contiguous points of equal elevation from the grid. The author used an on-the-shelf GIS package called Macgrizo [M1] to create contours for the Killeen Texas area. The second source of real data, which is the military counterpart to DEM data, is digital terrain elevation data (DTED), which currently is available in two resolutions: Level I, at 100 meter spacing, and Level II, at 30 meter spacing.

The simulated data is handcrafted by appealing to Macintosh Quickdraw graphics. A representative terrain containing four hillsides is depicted in the figure on the next page, where a query point is represented by the tip of the cursor (the arrowhead at the right). In this case, the partitioning algorithm during a preprocessing step created level one and level two cuts to segregate the four hillsides. The level three cuts and beyond partitioned each of the individual hills, using the nesting principle described earlier. Now comes execution time, and two-dimensional binary search. In this example, hills two and four were selected in the first binary cut, and hill two in the second cut. In the third cut, it was determined that the query point was not inside hill two's forty meter contour; in the fourth cut it was determined that the point was inside the twenty meter contour. Binary search concludes at this time because the contour containment graph has been exhausted. Therefore, the outer bracket is hill two's twenty meter contour, and the inner bracket is the forty meter contour. Associated with each of these two contours is the label "HILL2". The gradient computation deduces an easterly downhill slope; the slope is computed to be forty eight degrees; and the elevation interpolates to thirty three meters. Currently, the interpretation process says nothing about the relationship among HILL2 and the other hills; future work will address this issue.

Future Work.

The research to date has focused on a local interpretation of a query point. By definition, a local interpretation is limited to a description of a query point in terms of the label of the hill upon which it resides, the two contours which bracket it, an interpolated elevation, a slope value, and a directional gradient. This information is useful for localized reasoning about the immediate environs of a query point. A natural outgrowth of this work is to extend the reasoning to a more global capability. For example, one could utilize knowledge about the location of a hillside with respect to other hillsides in a specific region, to achieve context-cued line of sight reasoning or traversability planning.

To illustrate line-of-sight reasoning, consider the following example, based on the author's personal experience. In Grand Teton National Park in Wyoming, if one is on the western shore of Jenny Lake, the tallest visible peak is Teewinot Mountain, which looms spectacularly nearly a mile above the observer's head. One peak away is the Grand Teton, which although a thousand feet higher, may not even be seen from this vantage point because it is blocked from sight by Teewinot. The interpretation process described in Section IV would determine that the query point is on the eastern flank of Teewinot Mountain, on a moderate slope, at about 6600 feet elevation. Utilization of the directional gradient calculation would indicate that the direction to the tops of Teewinot and the Grand Teton are roughly the same, but the slope of the segment joining the query point to the top of Teewinot is greater than that drawn to the top of the Grand Teton. Hence, one concludes that line-of-sight westward to the Grand

is restricted by the intervening mass of Teewinot. Future work will involve refining and formalizing concepts such as these.

Already, the normal vector admissibility filtering technique has been extended to objects other than topographic contours. The Defense Mapping Agency produces a set of vector overlays corresponding to a transportation network, a hydrology network, obstacles, surface orientation, surface composition, and vegetation type. In addition, the DMA produces a gridded product called digital terrain elevation data (DTED), at both thirty and one hundred meter horizontal resolution. The vector products together with thirty-meter DTED in large part comprise what is known as tactical terrain data (TTD), a database being developed by DMA with the cooperation of the US Army Topographic Engineering Center [M2]. The integer-based decision rule derived from the law of cosines has proven to be of high utility in gauging proximity and inclusion with respect to the multi-megabyte vector databases contained in TTD.

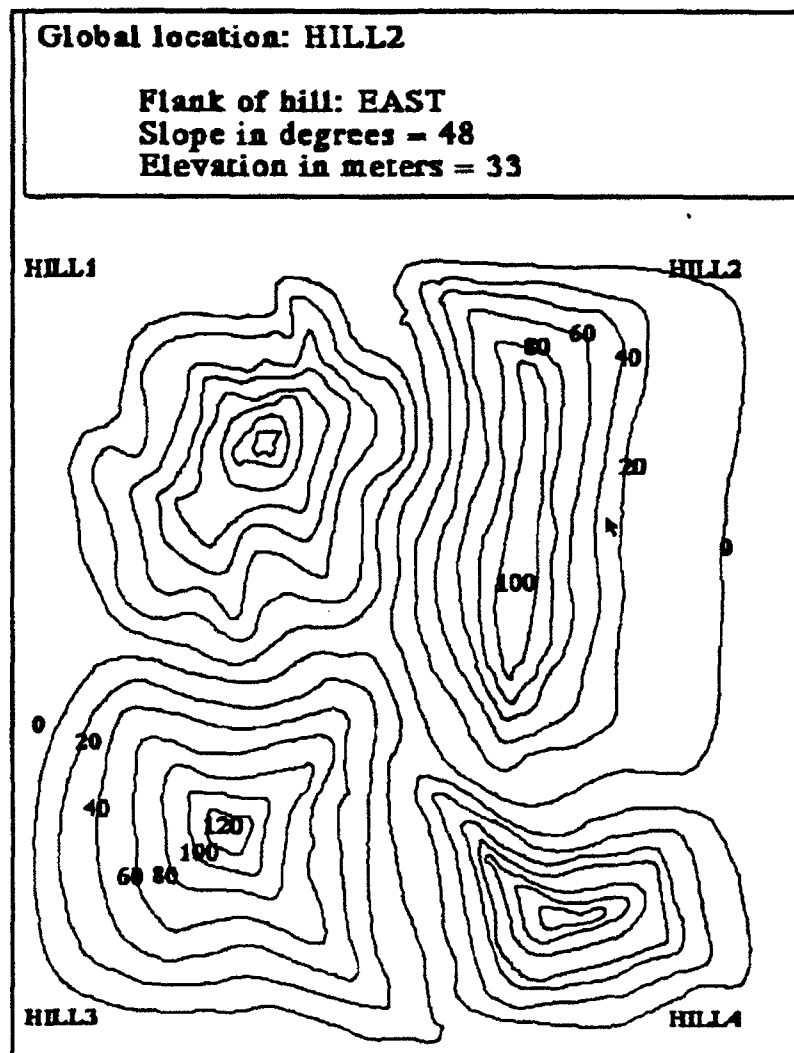


Figure 16. Interpreting a query point in terrain.

Conclusions.

Two-dimensional binary search has been utilized in conjunction with two new algorithms which avoid the expensive floating point operations associated with computing the normal vector, to produce an algorithm adept at locally interpreting topographic line maps. An interpretation consists of a human-like description of a query point in terms of its global location, interpolated elevation, local slope, and directional gradient. The search algorithm relies heavily upon proximity and inclusion algorithms developed with computational geometry research funded by the US Army. For credibility, the technique is being leveraged against multi-megabyte databases of contour information corresponding to actual terrain. An integer-based decision function which arbitrates when to drop the normal vector to an edge (during proximity testing) has proven to be extensible to objects other than elevation contours, such as segments of roads and streams, and polygons delimiting types of vegetation cover and surface material composition. As a byproduct of the research, an algorithm based on the ancient Cevian formula has been developed to find the nearest segment to a query point, without using any floating point operations whatsoever. New work will focus on extending the definition of map interpretation to be more globally descriptive of a terrain.

Bibliography

- [B1] Bentley, J., Programming Pearls, Addison-Wesley Publishing Company, Reading MA, 1986.
- [B2] Blum, H., A Transformation for Extracting New Descriptors of Shape, in *Symp. Models for Perception of Speech and Visual Form*, MIT Press, 1967.
- [C1] Cronin, T., Topographical Contour Betweenness Testing, US Army Signals Warfare Center Technical Report CSW-88-7, 1988.
- [C2] Cronin, T., Optimized Annulus-based Point-in-Polygon Inclusion Testing for d Dimensions, Transactions of the Seventh Army Conference on Applied Mathematics and Computing, West Point NY, 1989.
- [E1] Edelsbrunner, H., Algorithms in Combinatorial Geometry, Springer-Verlag, Berlin Germany, 1987.
- [F1] Fortune, S., A Sweepline Algorithm for Voronoi Diagrams, Proceedings of the Second Annual ACM Computational Geometry Symposium, 1986.
- [F2] Fortune, S., private communication, AT&T Bell Laboratories, March 1992.
- [K1] Kay, D., College Geometry, Holt, Rinehart, and Winston, Inc., New York NY, 1969.
- [K2] Khachiyan, L.G., A Polynomial Algorithm in Linear Programming, Soviet Math. Dokl., Vol 20, No.1, 1979.
- [K3] Kirkpatrick, D.G., Optimal Search in Planar Subdivisions, SIAM J. Comp. 12, 1983.
- [K4] Kjellstrom, B., Be Expert with Map and Compass, rev. ed., Charles Scribner's Sons, New York NY, 1967.
- [M1] Macgridzo: The Contour Mapping Program for the Macintosh, Users Manual for Version 3, Rockware Inc., Wheat Ridge CO, 1990.
- [M2] Messmore, J. and L. Fatale, Phase I Tactical Terrain Data (TTD) Prototype Evaluation, US Army Engineer Topographic Laboratories Technical Report ETL-SR-5C, Ft. Belvoir VA, December 1989.
- [M3] Mitchell, J., private communication, SUNY Stony Brook, July 1992.
- [T1] Tech-Tran, Vol. 13, Num. 4, US Army Engineer Topographic Laboratory, Ft. Belvoir VA, Fall 1988.

**CURRENT TRENDS OF RESEARCH IN STATISTICS
SMALL SAMPLE ASYMPTOTICS, RESAMPLING
TECHNIQUES AND ROBUSTNESS***

C. Radhakrishna Rao

Center for Multivariate Analysis
The Pennsylvania State University
University Park, PA 16802

ABSTRACT

In many statistical problems, it is often difficult to evaluate the exact distribution of a statistic. In such cases we resort to asymptotic methods. The paper surveys some recent results on asymptotic expansions of the distribution of a statistic, with successive terms providing improvement in accuracy but decreasing in magnitudes with orders of successive powers of $1/n$ or $1/\sqrt{n}$, where n is the sample size. It also describes the methods of jackknife and bootstrap for making approximate computations of some distributional aspects of estimators and test criteria. Recent work on robust inference is briefly mentioned.

Key Words: Berry-Esseen bound, Central limit theorem, Charlier differential series, Cornish-Fisher expansion, Edgeworth expansion, Fisher-Rao theorem, Hampel's expansion, Laplace expansion, Robust inference, Saddle point approximation, Second order efficiency, Small sample asymptotics.

AMS Classification Index: 62E20, 60F05

*Paper presented at the Tenth Army Conference on Applied Mathematics and Computing, 16-19 June 1992.

Research sponsored by the U.S. Army Research Office under Grant DAAL03-89-K-0139.

1. INTRODUCTION

1.1 Exploratory and Inferential Data Analysis

Broadly speaking, eliciting information (or drawing inference) from observed data is accomplished in two stages. The first is EDA (exploratory data analysis), the purpose of which is to understand the nature of the data and the underlying stochastic mechanism. This is done through a descriptive analysis of data and graphical displays, which can be of help in scrutinizing data for inhomogeneities, editing, faking and other inconsistencies and in suggesting a probability model underlying the data. The second is IDA (inferential data analysis) whose purpose is to test the appropriateness of the stochastic model selected for the analysis of observed data, to estimate the unknown parameters in the model, and, what is more important, to evaluate in quantitative terms the uncertainty in the conclusions drawn (inference) from the data. The two stages of data analysis, EDA and IDA, are represented in Table 1, with some additional explanations on the generation of the data and the end results of analysis.

1.2 Parametric and Nonparametric Models

In IDA, the analysis of data is dictated by the chosen (or specified) stochastic model and, therefore, the validity of our conclusions depends on how accurate the specification is. In the early stages of the development of statistical methodology, much emphasis was laid on parametric models, and especially the normal distribution was taken as basic to all quantitative data. Other more elaborate parametric models such as the Pearsonian system of frequency curves were introduced but mathematical difficulties prevented their full exploitation in applications. At this stage, attempts were made to develop nonparametric inference without assuming any particular family of probability distributions, i.e., under which the conclusions drawn remain valid whatever may be the underlying distribution. However, the inference in such a case is not as precise as it would be under a valid parametric model.

These two methodologies are at the extreme ends of our knowledge about the specification of the stochastic model. In the sixties some new concepts were introduced by combining the parametric and nonparametric inference procedures. It was assumed that a basic parametric model holds, but the observed data may be contaminated with extraneous observations. To accommodate such situations, the specification is enlarged by considering mixtures of the basic model with possible contaminating distributions. Methods are then developed to ensure robustness against the contaminating distributions. The different approaches to specification and the nature of inference associated with them are given in Table 2.

Table 1

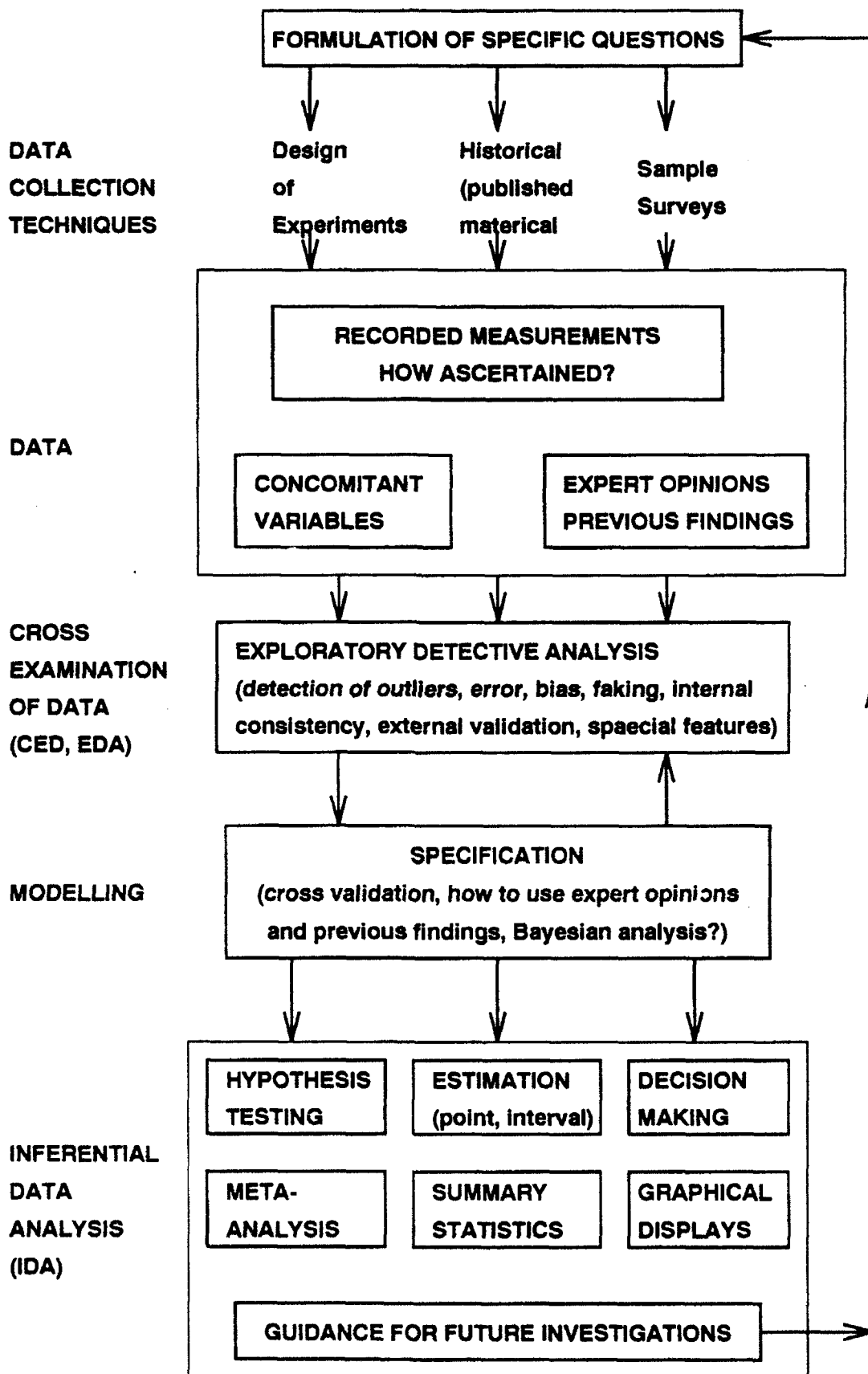


Table 2: Inferential Data Analysis

Specific parametric model	Parametric model with a neighborhood	Nonparametric model
$F(\theta)$ Fisher (1922)	$(1-\alpha)F(\theta) + \alpha G$ Huber (1964)	Wide class of distributions Pitman (1937)
<ul style="list-style-type: none"> • What is the appropriate method of data analysis for the chosen model? • Inference specific to the chosen model 	<ul style="list-style-type: none"> • For what class of models a particular method is appropriate or meaningful? • Inference on F, robust against G, and other possible errors in data. 	<ul style="list-style-type: none"> • How best to use ranked data? • Inference valid for any underlying distribution
Too much reliance on the chosen model. Inference could be misleading	Provides robustness or insurance against possible deviations from the assumed model	Generally does not result in efficient use of data

1.3 Asymptotic Theory

However, in IDA, after choosing a stochastic model, we are faced with the problem of computing the probability distributions of statistics used as estimators or test criteria, or in general of evaluating complicated integrals over specified regions of the sample space. This is not easy to do exactly and approximations have to be made. This gave rise to asymptotic theories of inference based on limit theorems which ensured the accuracy of certain approximations as the sample size increased. However, different approximations could not be compared on the basis of limit theorems alone, which necessitated investigations on the rates of convergence. New concepts such as second order efficiency (Rao (1961)), small sample (or higher order) asymptotics (Barndorff-Nielsen and Cox (1989), Field and Ronchetti (1990)) and resampling techniques such as Jackknife (Quenouille (1956), Tukey (1958)) and bootstrap (Efron (1982)) were introduced. Much of the current research is centered round these concepts, which are described in this paper. The basic results and techniques of asymptotic theory are presented in Table 3.

Table 3: Asymptotic Theory

Parametric		Non-parametric
Normal	Non-normal	
Exact sample theory (generally available)	<ul style="list-style-type: none"> • Central limit theorem • Berry-Esseen bound • Edgeworth expansion • Cornish-Fisher expansion • Saddle point approximation • Conjugate density and related techniques • Second (third) order efficiency 	<ul style="list-style-type: none"> • Central limit theorem • Empirical Edgeworth expansion
	Resampling techniques <ul style="list-style-type: none"> • Jackknife • Bootstrap 	

Note 1. In this paper, only the key references to original papers are given. For the numerous other relevant references, the papers and/or the text books in which they can be found are mentioned.

Note 2. The following abbreviations and notations are used.

df = distribution function, pd = probability distribution.

EE = Edgeworth expansion, Φ = df of $N(0,1)$

ϕ = pd of $N(0,1)$, $N(0,1)$ = standard normal variable

iid = independent and identically distributed

Field and Rochetti (1990) = FR, Barndorff-Nielsen and Cox (1980) = BC

Bhattacharya and Denker (1991) = BD

2. EDGEWORTH AND RELATED EXPANSIONS

2.1 Berry-Esseen Theorem

One of the first results in asymptotic theory is the central limit theorem which says that the df (distribution function) of the standardized average of a sample of n iid observations

tends to normal as $n \rightarrow \infty$. The following theorem gives the conditions under which this result holds and an expression to the upper bound for error can be found.

Theorem 2.1 (Berry-Esseen). Let X_1, \dots, X_n be n iid random variables (r.v.'s) with a common df F such that $E(X_i) = 0$, $E(X_i^2) = \sigma^2 > 0$, $E|X_i|^3 = \rho < \infty$. Denote by F_n the distribution of the standardized statistic $N^{-1/2}(X_1 + \dots + X_n)/\sigma$. Then, for all n

$$\sup_t |F_n(t) - \Phi(t)| \leq \frac{3\rho}{\sigma^3 \sqrt{n}} \quad (2.1.1)$$

where $\Phi(t)$ is the df of the standard normal variable.

The best known constant which replaces 3 in (2.1.1) is 0.7975. The result of Theorem 2.1 has been generalized to non iid r.v.'s, U-statistics, linear combination of ordered statistics, symmetric functions of iid r.v.'s and so on. For references to papers on the extensions and generalizations of Berry-Esseen theorem, the reader is referred to Field and Ronchetti (1990), which we indicate by FR in the sequel.

The Berry-Esseen theorem raises the following question. Under what conditions can we obtain a complete asymptotic expansion

$$F_n = \sum_{j=0}^{\infty} \frac{A_j(t)}{n^{j/2}}$$

such that

$$\left| F_n - \sum_{j=0}^k \frac{A_j(t)}{n^{j/2}} \right| \leq n^{-(k+1)/2} C_k(t) \quad (2.1.1)$$

We provide some attempts that have been made to solve the above problem, and the results that are now known.

2.2 Charlier Differential Series

Let $F(x)$ and $G(x)$ be two distribution functions such that all the derivatives of G vanish at the extremes of the range of x . Denote the cumulants of F and G by β_r and γ_r , $r = 1, 2, \dots$, respectively. Charlier established the following formal expansion

$$F(x) = \exp \left\{ \sum_{k=1}^{\infty} (\beta_r - \gamma_r) \frac{(-D)^r}{r!} \right\} G(x) \quad (2.2.1)$$

where D denotes the differential operator and $e^D = \sum_{j=0}^{\infty} (j!)^{-1} D^j$.

The series (2.2.1) known as Charlier differential series enables us to find the expansion for the df of an r.v. in terms of its cumulants and derivatives of any given df. An important application of this idea, which led to a number of important developments is due to Edgeworth.

2.3 Edgeworth Expansion

Edgeworth (1905) obtained Charlier type expansion for the distribution of the average of a sample of n observations in terms of the normal distribution and its derivatives. Modern versions of EE (Edgeworth expansion) and its extension to other statistics are as follows.

Theorem 2.2. Let X_1, \dots, X_n be iid r.v.'s with a common df F . Let

$$E(X_i) = 0, E(X_i^2) = \sigma^2 < \infty \text{ and } F_n(x) = [n^{1/2} \bar{X} / \sigma < x].$$

If F is not a lattice df, and if the third moment μ_3 of F exists, then uniformly in x

$$F_n(x) = \Phi(x) + \frac{1}{\sqrt{n}} \frac{\mu_3}{6\sigma^3} (1-x^2)\phi(x) + o(n^{-1/2}).$$

Theorem 2.3. Let X_1, \dots, X_n be iid r.v.'s with a common df F and characteristic function ψ . Further let

$$E(X_i) = 0, \quad E(X_i^2) = \sigma^2 < \infty$$

and $f_n(x)$ be the pd of $n^{1/2}\bar{X}/\sigma$. Suppose that the standardized moments $\lambda_i = \mu_i/\sigma^i$ exist for $i = 2, \dots, k$ and $|\psi|^v$ is integrable for some $v \geq 1$. Then f_n exists for $n \geq v$ and as $n \rightarrow \infty$

$$f_n(x) = \phi(x) + \phi(x) \sum_{r=3}^k P_r(x)/n^{(r/2)-1} + o\left(\frac{1}{n^{(k/2)-1}}\right)$$

uniformly in x . Here P_r is a real polynomial of degree $3(r-2)$ depending only on $\lambda_1, \dots, \lambda_r$ but not on n and k (or otherwise on F).

For instance, the first few successive approximations to the density function are

$$\begin{aligned} f_n(x) &= \phi(x) \\ f_n(x) &= \phi(x) \left\{ 1 + \frac{\lambda_3}{6\sqrt{n}}(x^3 - 3x) \right\} \\ f_n(x) &= \phi(x) \left\{ 1 + \frac{\lambda_3}{6\sqrt{n}}(x^3 - 3x) + \frac{\lambda_4}{24n}(x^4 - 6x^2 + 3) + \frac{\lambda_3^2}{72n}(x^6 - 15x^4 + 45x^2 - 15) \right\}. \end{aligned}$$

Note that when $\lambda_3 = 0$ (i.e., when F is symmetric), the normal approximation is accurate to order $(1/n)$.

EE has been extended to U statistics of degree 2 (Bickel, Götze and van Zwet), regression models (Qumsiyeh), maximum likelihood estimators (Skovgaard), M-estimators, likelihood ratios and other statistics (Bhattacharya and Denker (1990)), which we indicate by BD in the sequel, and functions of mean values of functions of vector random variables (Bhattacharya and Ghosh (1978)). The results of Bhattacharya and Ghosh have been recently extended by Bai and Rao (1991) to cases where some of the components of the vector variables are not continuous. References to the above authors are given in FR and BD.

2.4 Cornish-Fisher Expansion

Consider an estimator $\hat{\theta}_n$ of a real parameter θ lying in an open set Θ , and let s/\sqrt{n} be an estimate of its standard error. Suppose that $G_n(x;\theta)$, the df of $\sqrt{n}(\hat{\theta}_n - \theta)/s$ has the asymptotic expansion

$$G_n(x;\theta) = \Phi(x) + \sum_{r=1}^{k-2} n^{-r/2} q_r(x;\theta) \phi(x) + o(n^{-(k-2)/2}) = \Psi_{k,n}(x;\theta) + o(n^{-(k-2)/2}) \quad (2.4.1)$$

uniformly in $x \in \mathbb{R}^1$ and θ in any compact $K \subset \Theta$. In (2.4.1), $q_r(x;\theta)$ are polynomials in x whose coefficients are smooth functions of θ (say, $(k-2)$ times continuously differentiable in θ). For purposes of obtaining a confidence interval of θ or testing a hypothesis concerning θ , we need upper and lower tail percentage points of $G_n(x;\theta)$. We can use the expansion (2.4.1) to obtain a good approximation to the upper p -fractile by finding $x_{n,p}$ such that

$$G_n(x_{n,p};\theta) = p + o(n^{-(k-2)/2})$$

or equivalently

$$\Psi_{k,n}(x_{n,p};\theta) = p + o(n^{-(k-2)/2}). \quad (2.4.2)$$

The equation (2.4.2) can be solved recursively by starting with the initial approximation x_p , the upper p -fractile of the standard normal distribution in the form

$$x_{n,p} = x_p + \sum_{r=1}^j n^{-r/2} c_r(x_p;\theta) + o(n^{-j/2}), \quad 0 \leq j \leq k-2 \quad (2.4.3)$$

which is known as Cornish-Fisher expansion.

For instance

$$c_1(x_p; \theta) = -q(x_p; \theta), \quad c_2(x_p; \theta) = -2^{-1} x_p q_1^2(x_p; \theta) + q_1(x_p; \theta) q_1'(x_p; \theta) - q_2(x_p; \theta)$$

where q_1' is the derivative of q_1 .

For proofs and examples, the reader is referred to BD.

2.5 Laplace Approximation

Consider an integral of form

$$g_n = \int_a^b e^{-nr(y)} h(y) dy. \quad (2.5.1)$$

Assume that the minimum of $r(y)$ is attained at $\hat{y} \in (a, b)$, that $r'(\hat{y}) = 0$ and $r''(\hat{y}) > 0$ and that $h(\hat{y}) \neq 0$. Then, we can expand $e^{-nr(y)}$ at \hat{y} and approximate (2.5.1) by omitting terms involving higher derivatives of $r(y)$

$$\begin{aligned} g_n &\doteq e^{-nr(\hat{y})} \sqrt{\frac{2\pi}{nr''(\hat{y})}} \int_{-\infty}^{\infty} [h(\hat{y}) + (y-\hat{y})h'(\hat{y}) + \dots] \phi(y-\hat{y}, [nr''(\hat{y})]^{-1}) dy \\ &\doteq e^{-nr(\hat{y})} h(\hat{y}) \left[\frac{2\pi}{nr''(\hat{y})} \right]^{1/2} \{1 + O(n^{-1})\}. \end{aligned} \quad (2.5.2)$$

Suppose that $r(y)$ is minimized at a and $r'(y) \neq 0$ there. Then

$$g_n = e^{-nr(a)} \left\{ \frac{h(a)}{nr'(a)} + O(n^{-2}) \right\} \quad (2.5.3)$$

with a similar result if the $r(y)$ is minimized at b . The multiparameter version of (2.5.2) is

$$g_n = \int_D e^{-nr(y)} h(y) dy = \frac{e^{-nr(\hat{y})} h(\hat{y}) (2\pi)^{m/2}}{[n|r''(\hat{y})|]^{1/2}} \{1 + O(n^{-1})\} \quad (2.5.4)$$

where it is assumed that $r(y)$ takes its minimum at y in the interior of $D \subset \mathbb{R}^m$, the gradient is zero and the Hessian $\gamma''(y)$ is positive definite.

The Laplace approximations (2.5.2-2.5.4) have proved very useful in Bayesian inference. A series of papers by Tierney, Kass and Kadane give a number of applications involving the evaluation of posterior expectation of a parametric function. For further details regarding Laplace approximation and references to the above authors, the reader is referred to Barndorff-Nielsen and Cox (1989, Ch. 3), which will be denoted by BC in the sequel, and Reid (1991).

2.6 Stochastic Expansions

Sometimes, it is useful to obtain a stochastic expansion of a random variable in terms of other variables, which are easy to handle. To define a stochastic expansion, consider a base set of sequences $\{b_{hn}\}$, $h = 0, 1, \dots$ such that as $n \rightarrow \infty$, $b_{hn} = o(b_{h-1,n})$ with $b_{0n} = 1$. Typical examples are

$$\begin{aligned} b_{0n} &= 1, b_{1n} = 1/\sqrt{n}, b_{2n} = 1/n, \dots \\ b_{0n} &= 1, b_{1n} = 1/n, b_{2n} = 1/n^2, \dots \end{aligned} \quad (2.6.1)$$

Suppose that $\{Y_n\}$ is a sequence of continuous r.v.'s such that

$$Y_n = X_0 + X_1 b_{1n} + \dots + X_h b_{hn} + O_p(b_{h+1,n}) \quad (2.6.2)$$

where $\{X_0, X_1, \dots\}$ have a distribution not depending on n . We call an expansion of the form (2.6.2) a stochastic expansion of Y_n . It is seen that as $n \rightarrow \infty$, the distribution of Y_n converges to that of X_0 .

For example, it is shown in Cox and Reid (1987) that if Y_n follows a chisquare distribution with n degrees of freedom, then

$$(Y_n - n)(2n)^{-1/2} = X_0 + \frac{\sqrt{2}}{3n^{1/2}}(X_0^2 - 1) + O_p(n^{-1}) \quad (2.6.3)$$

where X_0 is a standard normal variable. Under some conditions, stochastic expansions and asymptotic expansions for distribution functions are equivalent. If

$$Y_n = X_0 + n^{-1/2}X_1 + O_p(n^{-1}) \quad (2.6.4)$$

then, as demonstrated in Cox and Reid (1987), $F_n(y)$, the df of Y_n has the form

$$F_n(y) = F_0(y)\{1 + n^{-1/2}a_1(y) + O(n^{-1})\} \quad (2.6.5)$$

where

$$a_1(y) = -E(X_1|X_0=y)f_0(y)[F_0(y)]^{-1}.$$

and $E(X_1|X_0)$ is the conditional mean. If $E(X_1|X_0) = 0$, then (2.6.5) reduces to $F_0(y)\{1 + O(n^{-1})\}$.

Stochastic expansions of statistics are useful in deriving stochastic expansions of their df's. For instance, in M-estimation (including maximum likelihood) through an estimating equation, it is often possible to obtain a stochastic expansion of the estimator. This enables the derivation of distributional results concerning the estimator. For details, the reader is referred to BC, BD and Cox and Reid (1987).

3. SECOND ORDER EFFICIENCY

Following the work of Fisher (1925), Rao (1961) introduced the concept of second order efficiency (SOE) of an estimator. [In modern terminology, my SOE is called TOE (third order efficiency)]. If $I_n(\theta)$ is Fisher information on θ in a sample of size n and $I_{t_n}(\theta)$ is the corresponding information in a Fisher consistent estimator t_n , Fisher suggested the expression

$$e = \lim_{n \rightarrow \infty} (I_n(\theta) - I_{t_n}(\theta)) \quad (3.1)$$

as the limiting loss of information. This criterion is difficult to work with. Rao (1961) defined t_n to be first order efficient (FOE) if

$$\sqrt{n}(n^{-1}\ell_n(x,\theta) - \beta[t_n(x) - \theta]) \rightarrow 0 \quad \text{in } P_\theta \text{ probability as } n \rightarrow \infty \quad (3.2)$$

where $\ell_n(x,\theta)$ is the score (derivative of log likelihood based on the sample). The second order efficiency is defined to be the minimum asymptotic variance of

$$\ell_n(x,\theta) - n\beta[t_n(x) - \theta] - \lambda n[t_n(x) - \theta]^2 \quad (3.3)$$

for suitable choices of β and λ . Under some conditions, it was shown (Rao, 1961) that the minimum asymptotic variance of (3.3) is

$$i(\theta) \left\{ \frac{\mu_{03} - 2\mu_{21} + \mu_{40}}{[i(\theta)]^2} - 1 - \frac{\mu_{11}^2 \mu_{30}^2 - 2\mu_{11}\mu_{30}}{[i(\theta)]^3} \right\} \quad (3.4)$$

where $i(\theta)$ is Fisher information in a single observation and

$$\mu_{rs} = E \left(\frac{1}{f} \frac{df}{d\theta} \right)^r \left(\frac{1}{f} \frac{d^2 f}{d\theta^2} \right)^s \quad (3.5)$$

where f is the pd of a single observation. It turned out that Fisher's evaluation of (3.1) was the same as (3.4). The conditions for equivalence of the two definitions remain to be worked out in a rigorous manner. The result (3.4) is termed as Fisher-Rao theorem by Efron (1975).

In a later paper, Rao (1962) obtained the approximate expression

$$E(\hat{t}_n - \theta)^2 \doteq \frac{A_1}{n} + \frac{A_2}{n^2} \quad (3.5)$$

when \hat{t}_n is suitably truncated and corrected for bias up to order $(1/n)$. The minimum value of $A_1 = 1/i$ and of A_2 is (3.4), which provides a decision theoretic interpretation of SOE.

Rao's work on SOE was extended by Pfanzagl, Ghosh and Subramanyam, Efron, and Akahira and Takeuchi. Some of these authors used the criteria of median unbiasedness and concentration around the true value. References to these authors and proofs of certain statements can be found in BD and in a forthcoming book by J. K. Ghosh (1993).

4. SADDLE POINT APPROXIMATION

A finite order EE provides a good representation of the df in the middle region, but not in the tails. Evaluation of tail probabilities is often needed in statistical inference. Use of EE for this purpose gives poor and sometimes negative results. A promising approach in such cases is the saddle point approximation introduced by Daniels (1954) originally for approximating the density function of \bar{x} , the average of n iid r.v.'s, x_1, \dots, x_n . We explain the technique for a general statistic $V_n = V_n(x_1, \dots, x_n)$ with density $f_n(v)$.

Let $M_n(t)$ and $K_n(t) = \log M_n(t)$ be the moment and cumulant generating functions of V_n and $R_n(t) = n^{-1} K_n(nt)$. Further let, for given v , \hat{t} be the unique value such that $R_n^{(1)}(\hat{t}) = v$, where $R_n^{(s)}$ denotes the s -th derivative of R_n . Then the density of V_n at v is

$$\begin{aligned} f_n(v) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} M_n(it) e^{-itv} dt \\ &= \frac{n}{2\pi i} \int_{c-i\infty}^{c+i\infty} e^{n[R_n(t)-tv]} dt. \end{aligned} \quad (4.1)$$

where c is any real number in the interval where the moment generating function exists. Expanding at $t = \hat{t}$,

$$n[R_n(t)-tv] = n[\hat{R}_n - \hat{t}v] + n\hat{R}_n^{(2)} \frac{(\hat{t}-t)^2}{2!} + \dots, \quad (4.2)$$

substituting in (4.1) and integrating term by term, we obtain

$$f_n(v) = \hat{f}_n(v) \left\{ 1 + \frac{1}{n} \left(\frac{\rho_{4n}}{8} - \frac{5\rho_{3n}^2}{24} \right) + \dots \right\} \quad (4.3)$$

where

$$\hat{f}_n(v) = \left(\frac{n}{2\pi \hat{K}_n^{(2)}} \right)^{1/2} e^{n(\hat{K}_n - i v)}, \quad \rho_{rn} = \rho_{rn}(\hat{t}) = \hat{K}_n^{(r)} / (\hat{K}_n^{(2)})^{r/2}.$$

The expression (4.3) will be an asymptotic expansion if ρ_{rn} , $r \geq 3$ are of order at most $O(1)$ so that the first term dropped is $o(\text{the last term included})$.

If $V_n = \bar{x}$, we obtain the expression given by Daniels (1954)

$$f_n(\bar{x}) = \left[\frac{n}{2\pi K^{(2)}(\hat{t})} \right]^{1/2} e^{n[K(\hat{t}) - i \bar{x}]} \times \left\{ 1 + \frac{3\rho_4(\hat{t}) - 5\rho_3^2(\hat{t})}{24n} + O(n^{-2}) \right\} \quad (4.4)$$

where $K(t)$ is the cumulant generating function of x_i , an individual observation, $K^{(1)}(\hat{t}) = \bar{x}$, and $\rho_n = K^{(r)}(\hat{t}) / [K^{(2)}(\hat{t})]^{r/2}$. It is customary to use as the approximation to $f_n(\bar{x})$, a normalized version of (4.4)

$$f_n(\bar{x}) = c \left[\frac{n}{K^{(2)}(\hat{t})} \right]^{1/2} e^{n[K(\hat{t}) - i \bar{x}]} \{ 1 + O(n^{-1}) \} \quad (4.5)$$

where c is determined so that the right hand side of (4.5) integrates to unity up to $o(n^{-1})$.

Expressions such as (4.3-4.5) are obtained by using methods of deepest descent for evaluating the complex integrals involved or the method of exponential tilting or using the

conjugate density approach. The second approach involves the use of EE of the pd of \bar{x} from the conjugate density $f(y)\exp[\hat{t}y-K(\hat{t})]$ with mean $E(y) = \bar{x}$.

The saddle point approximations have now been worked out for a number of statistics. For further details, the reader is referred to FR and a thesis by Yau (1989).

5. HAMPEL'S TECHNIQUE

Consider the problem of approximating f_n , the p.d. of \bar{x} , the mean of n iid r.v.'s with a common p.d. f . The conjugate density of f centered at t is

$$h_t(x) = c(t)f(x)e^{\alpha(t)(x-t)} \quad (5.1)$$

where $\alpha(t)$ is the solution of

$$\int (x-t)h_t(x)dx = 0. \quad (5.2)$$

The existence of $\alpha(t)$ and its derivatives up to order 4 is ensured if we assume

$$\int x^r e^{\beta x} f(x) dx < \infty \text{ for } r \text{ up to } 5.$$

Then

$$f_n(t) = nc^{-n(t)} \int \dots \int h_t(nt - \sum_{i=1}^{n-1} x_i) \prod_{i=1}^{n-1} h_t(x_i) dx_1 \dots dx_{n-1} = c^{-n(t)} h_{t,n}(t) \quad (5.3)$$

where $h_{t,n}(t)$ is the density of \bar{x} with underlying density $h_t(x)$. Recall that

$$E_{h_t}(\bar{x}) = t, \quad V_{h_t}(\bar{x}) = \sigma^2(t)/n. \quad (5.4)$$

Hence $h_{t,n}(t)$ can be approximated by $n^{1/2}/\sqrt{2\pi} \sigma(t)$, and $h'_{t,n}(t)/h_{t,n}(t)$ by $\sigma'(t)/\sigma(t)$, each

with errors of order $1/n$. The term $n^{-1/2}$ disappears since we are evaluating the density at the mean. From this

$$\frac{f'_n(t)}{f_n(t)} = -n \frac{c'(t)}{c(t)} - \frac{\sigma'(t)}{\sigma(t)} + O(n^{-1}) = -n\alpha(t) - \frac{\sigma'(t)}{\sigma(t)} + O(n^{-1}). \quad (5.5)$$

Using (5.5), $f_n(t)$ is obtained by integration, which may have to be done numerically. In FR, the approximation (5.5) for f'_n/f_n is shown to be extremely good compared to others.

6. EVALUATION OF TAIL PROBABILITIES

Tail probability approximation is another area of great development. One method is to use a good approximation to the pd of a statistic and obtain the tail area by numerical integration. Another is to use the saddle point approximation directly to the integral representing the tail probability. A third method is to evaluate the integral for the tail probability directly by numerical integration.

Consider the example given in FR from the paper by Helstrom and Ritcey (1984). The problem is to evaluate the tail probability of $T_n > x_0$ where

$$T_n = \frac{1}{2} \sum_{j=1}^n (x_j^2 + y_j^2) \quad (6.1)$$

and $x_j \sim N(s_j, 1)$, $y_j \sim N(t_j, 1)$ are iid r.v.'s. It is seen that under the non-null hypothesis, the moment generating function of T_n is

$$M_n(t) = (1-t)^{-n} e^{st/(1-t)} \quad (6.2)$$

when $s = (\sum s_j^2 + \sum t_j^2)/2$. If s_j and t_j are themselves r.v.'s with $M_n^{(S)}$ as the moment generating function of S , then

$$M_n(t) = (1-t)^{-n} M_n^{(S)}[-t/(1-t)]. \quad (6.3)$$

The pd of T_n is

$$f_n(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} M_n(it) e^{-ix} dt \quad (6.4)$$

and the tail probability is

$$\int_{x_0}^{\infty} f_n(x) dx = \frac{1}{2\pi i} \int_{\tau-i\infty}^{\tau+i\infty} M_n(t) e^{-tx_0} dt = \int_{\tau-i\infty}^{\tau+i\infty} e^{w_n(t)} dt \quad (6.5)$$

where $w_n(t) = -\log t - tx_0 + K_n(t)$, $K_n = \log M_n$. We have two alternatives; either to use saddle point approximation by expanding the integrand in (6.5) around the saddle point and use the method of steepest descent or to apply numerical integration directly to (6.5) for evaluating the integral along the contour defined in the (x,y) plane by $\text{Imag}[x+iy] = 0$. It is found that the latter approach gave slightly better results. Perhaps, with the development of accurate numerical integration methods, the need for asymptotic expansions will become less important.

For further details on the subject and attempts at finding highly accurate explicit approximation formulas for tail probabilities stemming from the work of Luganani and Rice (1980), the reader is referred to FR and Yau (1989).

7. JACKKNIFE

Let x_1, \dots, x_n be iid r.v.'s having a common df F , and $\hat{\theta}$ be an estimate of a parameter θ . As a first step, one would like to know the properties of $\hat{\theta}$ in terms of bias and variance.

The Jackknife methodology assumes that

$$E(\hat{\theta} - \theta) = \frac{a_1}{n} + \frac{a_2}{n^2} + \dots, \quad E(\hat{\theta}_1 - \theta) = \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \dots \quad (7.1)$$

where $\hat{\theta}_{-i}$ is the estimator based on all the observations except x_i . The expressions

$$\hat{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{-i}, \quad i = 1, \dots, n \quad (7.2)$$

are called pseudvalues, and their average value

$$\hat{\theta}_J = n\hat{\theta} - (n-1)\hat{\theta}_\bullet \quad (7.3)$$

where $\hat{\theta}_\bullet = (\hat{\theta}_{-1} + \dots + \hat{\theta}_{-n})/n$, is called the Jackknife estimator of θ . It is seen that the bias in $\hat{\theta}_J$ is of the order of $1/n^2$.

Tukey (1958) suggested an estimate of the variance of $\hat{\theta}_J$ in the form

$$\hat{V}(\hat{\theta}_J) = [n(n-1)]^{-1} \sum_1^n (\hat{\theta}_i - \hat{\theta}_J)^2 = (n-1)n^{-1} \sum_1^n (\hat{\theta}_{-i} - \hat{\theta}_\bullet)^2. \quad (7.4)$$

and conjectured that the statistic

$$\frac{\hat{\theta}_J - \theta}{[\hat{V}(\hat{\theta}_J)]^{1/2}} \quad (7.5)$$

has approximately a t distribution with $(n-1)$ degrees of freedom. The results (7.4) and (7.5) may not be valid in all cases, but are found to provide good approximations in cases of regular estimators which are asymptotically locally linear.

More generally, we can delete d observations at a time and compute the estimate based on the remaining set of observations. Let $\hat{\theta}_s$ be such an estimate for a particular deletion (s) and $\hat{\theta}$ be the estimate obtained from all the observations. Then an estimate of $V(\hat{\theta})$ is

$$\frac{d!(n-d)!r}{n!d} \sum_s (\hat{\theta}_s - \hat{\theta})^2 \quad (7.5)$$

where the summation is over all possible deletions of d observations and $r = n - d$. When $\hat{\theta}$ is not a sufficiently smooth function, then the expression (7.5) with a large d is a better estimator of $V(\hat{\theta})$ than (7.4) based on $d = 1$.

Further details and references to the key contributors to Jackknife beginning with the pioneering contribution by Quenouille (1955) can be found in a survey paper by Peddada (1993).

8. BOOTSTRAP

Bootstrap is another tool like jackknife for studying the properties of estimators. It is more general than jackknife and can handle problems where jackknife fails.

Let θ be a parameter of a df F and $t_n = t(x_1, \dots, x_n)$ be an estimator of θ based on iid r.v.'s, x_1, \dots, x_n from F . We wish to know the characteristics of t_n such as the following, under repeated sampling from F .

$$\text{Bias: } E_F(t_n) - \theta = b. \quad (8.1)$$

$$\text{Variance: } E_F(t_n^2) - [E_F(t_n)]^2 = v. \quad (8.2)$$

$$\text{Tail area: } P_F(t_n \geq y) = p_y. \quad (8.3)$$

The actual evaluation of these quantities is extremely complicated even under simple forms for F . Bootstrap is a general resampling technique to obtain their approximate values.

The first step in bootstrap methodology is the estimation of F based on the sample. In the nonparametric situation F is simply estimated by $\hat{F}_n = F_n$, the empirical df. In the parametric situation, if $F(x) = F(x; \theta)$ where F has a known form but θ is an unknown parameter, then an estimate of F is $\hat{F}_n = F(x; \hat{\theta})$ where $\hat{\theta}$ is an efficient estimate of θ .

We now consider \hat{F}_n as our population and denote by (x_1^*, \dots, x_n^*) , n iid observations from \hat{F}_n . Define $t_n^* = t(x_1^*, \dots, x_n^*)$ and

$$E_{\hat{F}_n} t = b^* \quad (8.4)$$

$$E_{\hat{F}_n} (t_n^*)^2 - [E_{\hat{F}_n} (t_n^*)]^2 = v^* \quad (8.5)$$

$$P_{\hat{F}_n} (t_n^* \geq y) = p_y^* \quad (8.6)$$

Since \hat{F}_n is completely known, the quantities (8.4)-(8.6) can in principle be computed, although it may not be easy.

Efron (1975) suggested approximating them by simulation. That is, we repeatedly draw samples (x_1^*, \dots, x_n^*) from \hat{F}_n and compute t^* for each sample. Suppose that we have drawn B samples, in which case we have B values of t^*

$$t_1^*, \dots, t_B^* \quad (8.7)$$

Then b^* and v^* are estimated by the average and variance of the values in (8.7). We represent them by \hat{b}^* and \hat{v}^* respectively. Similarly we can find the proportion of values in (8.7) which are equal to or greater than y , which is represented by \hat{p}_y^* . It is claimed that when B is sufficiently large,

$$\hat{b}^* \sim b, \quad \hat{v}^* \sim v, \quad \hat{p}_y^* \sim p_y \quad (8.8)$$

The claim is substantiated by empirical studies and theoretical investigations on the rates of convergence of the bootstrap estimates b^* , v^* and p_y^* to the true values b , v and p_y respectively.

The bootstrap methodology can also be used to find the null distribution of certain test criteria. For instance, let the test criterion for testing that the mean of F has a specified value

μ_0 be $t = \sqrt{n}(\bar{x} - \mu_0)/s$, where s is the standard deviation computed from the sample. We need the distribution of t , or certain percentile points of its distribution.

Let \hat{F}_n be an estimate of F and x_1^*, \dots, x_n^* be a sample from \hat{F}_n . Further, let \bar{x}^* and s^* be the average and standard deviation of x_1^*, \dots, x_n^* . Construct the statistic $t^* = \sqrt{n}(\bar{x}^* - \bar{x})/s$ where \bar{x} is the average of the original sample. We obtain the distribution of t^* by drawing repeated samples from \hat{F}_n . Then

$$P_{\hat{F}_n}(t^* \geq y) \sim P_F(t \geq y). \quad (8.9)$$

Further details and references to numerous key papers on bootstrap can be found in a survey paper by Babu and Rao (1993).

9. ROBUST INFERENCE

It has been known for some time that certain statistical procedures are sensitive to slight departures from the assumptions on which they are based. But it is only in the sixties and seventies, systematic attempts were made to develop robust techniques. One of the main concerns was the effect of outliers, gross errors and contaminating observations on estimators. To study this, new concepts such as influence function, breakdown point and qualitative robustness were introduced.

As an alternative to the least squares method for the estimation of parameters in a linear model, which is unduly influenced by outliers, new methods of estimation known as M , L and R were introduced. The most popular among them is M -estimation based on a suitably chosen estimating equation. For example, to estimate a location parameter θ , an equation of the type

$$\sum_{i=1}^n \psi\left(\frac{x_i - \theta}{s}\right) = 0 \quad (9.1)$$

is used, where x_1, \dots, x_n are iid observations, θ is an unknown location parameter and s is some estimate of the scale parameter. To reduce the effect of outliers, ψ functions of the following forms shown in the accompanying diagram, are recommended instead of the endless line passing through the origin at 45° , which is the ψ function for least square estimation.

Figure 1

REDESCENDING M -ESTIMATORS

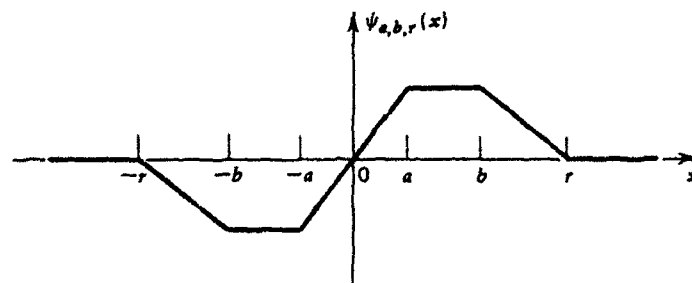


Figure 1. Shape of the Ψ -function (and the influence function) of Hampel's three-part redescending M -estimator.

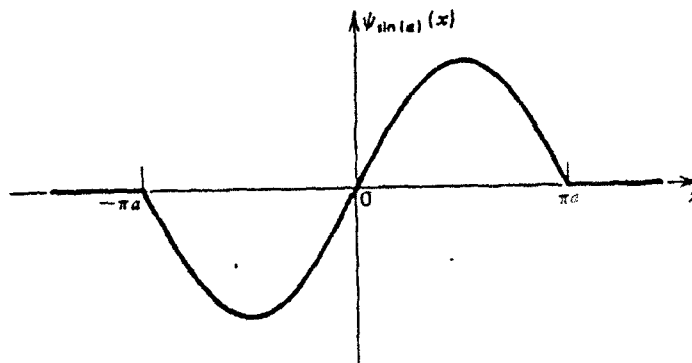


Figure 2. Andrew's sine function.

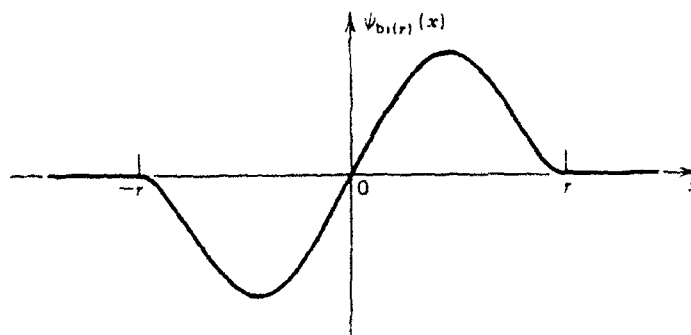


Figure 3. Tukey's biweight.

For details concerning robust inference, the reader is referred to the thesis of Ronchetti (1982) and the book by Hampel, Ronchetti, Rousseeuw and Stahel (1986). We can also consider M-estimation by introducing a loss function $\xi(x_i - \theta)$ and determining θ such that

$$\sum_{i=1}^n \xi(x_i - \theta) \quad (9.2)$$

is a minimum. In recent papers, Bai, Wu and Rao (1992) and Bai, Liu and Rao (1992) considered ξ functions of two types. One is the difference of two convex functions and another is non-increasing in the range $(-\infty, 0)$ and non-decreasing in the range $(0, \infty)$. With such loss functions, which cover a wide variety of shapes, an asymptotic theory is developed with a minimum number of assumptions.

Note 1. Asymptotitis: In the initial stages of development of statistics, much emphasis was laid on small sample theory, an exact treatment of which was possible under the normal distribution. Most of the current research papers deal with nonparametric or semiparametric situations and present asymptotic results. No attempt is made to examine how useful and relevant these results are for application in actual practice where we meet with only finite (small) samples. This phenomenon is aptly described by Tukey (1993) as *Asymptotitis*. Perhaps, we should be addressing ourselves to questions like: What sample sizes suffice to make asymptotic theory useful (or even relevant)?

Note 2. Most of the expansions considered in the literature use the limit distribution of the statistic under consideration as the leading term. There are other possibilities like the one considered by Rao (1951), which are likely to give improved results. Further research in this direction would be useful.

10. REFERENCES

1. Babu, G. J. and Rao, C. R. (1993). Bootstrap methodology, *Handbook of Statistics: Computational Statistics* (Ed. C. R. Rao), North-Holland.
2. Bai, Z. D. and Rao, C. R. (1991). Edgeworth expansion of sample means. *Ann. Statist.* 19, 1295-1315.
3. Bai, Z. D., Liu, Z. J. and Rao, C. R. (1993). On the strong consistency of M-estimates in linear models under a general discrepancy function, *Handbook of Statistics: Econometrics* (Eds. Maddala, Vinod and Rao), North-Holland.

4. Bai, Z. D., Rao, C R. and Wu, Y. (1992). A note on M-estimation of multivariate linear regression. Technical Report 92-14, CMA, Penn State.
5. Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for use in Statistics*. Chapman and Hall, London.
6. Bhattacharya, R. and Denker, M. (1990). *Asymptotic Statistics*. Birkhäuser Verlag, Basel.
7. Cox, D. R. and Reid, N. (1987). Approximations to noncentral distributions. *Canad. J. Statist.* 15, 105-114.
8. Daniels, H. E. (1954). Saddle point approximations in statistics. *Ann. Statist.* 26, 631-650.
9. Efron, B. (1975). Defining the curvature of a statistical problem (with application to second order efficiency) (with discussion). *Ann. Statist.* 3, 1189-1242.
10. Field, C. and Ronchetti, E. (1990). Small Sample Asymptotics, IMS Lecture Notes, Volume 13.
11. Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. A* 222, 309-368.
12. Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Philos. Soc.* 22, 700-725.
13. Ghosh, J. K. (1993). *Higher Order Asymptotics*.
14. Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics*. Wiley, New York.
15. Helstrom, C. W. and Ritcey, J. A. (1984). Evaluating radar detection probabilities by steepest descent integration. *IEEE Transactions on Aerospace and Electronic Systems*, AES-20, 5, 624-634.
16. Huber, P. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* 35, 73-101
17. Luganani, R. and Rice, S. O. (1984). Distribution of the ratio of quadratic forms in normal variables-numerical methods. *SIAM J. Scientific and Statistical Computing* 5, 476-488.
18. Peddada, S. D. (1993). Jackknife variance estimation and bias reduction. *Handbook of Statistics: Computational Statistics* (Ed. C. R. Rao). North-Holland.
19. Pitman, E.J.G. (1937). Significance test which can be applied to samples for any population. *J. Roy. Statist. Soc. (suppl)* 4, 119-130.

20. Quenouille, M. (1956). Notes on bias estimation. *Biometrika* 43, 353-360.
21. Rao, C. R. (1951). An asymptotic expansion of the distribution of Wilk's criterion. *Bull. Inter. Statist. Inst.* 33, 177-180.
22. Rao, C. R. (1961). Asymptotic efficiency and limiting information. *Proc. Fourth. Berkeley Symp. Math. Statist. Probability 1*, 531-546.
23. Rao, C. R. (1962). Efficient estimates and optimum inference procedures in large samples. *J. Roy. Statist. Soc. Ser B* 24, 46-63.
24. Reid, N. (1991). Approximations and asymptotics. In *Statistical Theory and Modelling* (Eds. Hinkley, Reid and Snell), Chapman and Hall, London 287-305.
25. Ronchetti, E. (1982). Robust Testing in Linear Models: The Infinitesimal Approach. Ph.D. Thesis submitted to the Swiss Federal Institute of Technology, Zurich.
26. Tukey, J. W. (1958). Bias and confidence in not quite large samples (abstract). *Ann. Math. Statist.* 29, 614.
27. Tukey, J. W. (1993). The major challenges for multiple-response (and multiple-adjustment) analysis. In *Multivariate Analysis: Future Directions* (Ed. C. R. Rao, North-Holland.
28. Yan, W. K. (1989). Saddle point approximations to the tail probabilities of the general statistics. Ph.D. Thesis submitted to the University of Toronto.

VARIATIONAL THEORY OF MOTION OF CURVED, TWISTED AND EXTENSIBLE ELASTIC RODS

Iradj Tadjbakhsh
Department of Civil & Environmental Engineering
Rensselaer Polytechnic Institute
Troy New York 12180-3590

and
Dimitris C. Lagoudas
Department of Aerospace Engineering
Texas A&M University
College Station, Texas, 77843-3141

The variational theory of three dimensional motion of curved twisted and extensible elastic rods is obtained based entirely on the kinematical variables of position and rotations. The constitutive relations that define the resistive couples and the axial force as gradients of the strain energy function are established. A candidate for the strain energy function, derived on the basis of classical assumptions, is presented.

INTRODUCTION

In the Historical Introduction of the "A Treatise on the Mathematical Theory of Elasticity," Love (1892) narrates that in 1742 Daniel Bernoulli wrote to Euler suggesting that the differential equation of the elastica could be found by making the integral of the work done or the square of the curvature a minimum. Acting on this suggestion Euler was able to obtain the differential equation of the elastica and the various forms of it. Thus the concept of the strain energy was born and the foundation of the variational theory of elastic rods were laid out. The equilibrium equations that were much later developed by Love are applicable to an initially bent and twisted rod.

Our aim in this paper is to establish a variational formulation for the title problem and in the process infer the existence of the strain energy function and determine the constitutive relations that relate this function to the bending and twisting couples and the axial force within the rod. This development together with equations of motion and the geometry of deformation define a direct approach and an exact nonlinear theory for the three dimensional motion of a one dimensional elastic medium capable of resisting bending twisting and extension. Going a step further, in order to actually construct an explicit form for the strain energy function, we enter the realm of hypothesis and use Kirchhoff's description of deformations in a thin rod. This view enables us to determine a strain energy function that can be used in engineering applications.

The recent history of investigations of the rod theories consists of developments along two separate streams, the direct approach and approximations from three-dimensional continuum. In the direct approach a one-dimensional continuum view is pursued and the medium is supposed endowed, in addition to its position, with vector fields, the directors,

that are to be interpreted appropriately to define bending, twist and extension properties of a rod. This approach has its origin in the work of E. and F. Cosserat (1909) and numerous investigations have contributed to it among them Naghdi (1982), Naghdi and Rubin (1984), Whitman and DeSilva (1970), Green and Laws (1966), and Eriksen (1970). Extensive investigation into the qualitative aspects of the nonlinear theory such as questions of existence of solutions and global behavior have been carried out by Antman (1976). His basic work entitled "The Theory of Rods" (1972) describes these theories both as approximations to the three-dimensional continuum theory and as a one-dimensional continuum with directors. The work presented here, although pertains to a one-dimensional continuum does not use directors, but is formulated entirely on the basis of kinematical quantities consisting of the position vector of points along the curve of centroids and the orientation angles of the cross sections of the rod relative to a fixed coordinate system. It is a generalization of the work of Tadjbakhsh (1966) in which the theory of planar motion of the extensible elastica was described.

The history of construction of approximate theories in the context of three-dimensional nonlinear continuum theory is also varied and to it many investigators including some of the above authors have contributed, see for example, Naghdi and Wenner (1974).

KINEMATICS

An elastica is a nearly uniform slender rod of finite length. In the unstressed state the centroids of the cross section form a space curve C that is called the reference curve with an arc length s . The orientation of the principal axes of the cross section vary continuously along the rod. This means that in the unstrained state the rod has arbitrary twist and curvatures. With respect to an inertial Cartesian frame \mathbf{x} the position of a point s in the unstrained state is denoted by $\mathbf{X} = X_i(s)\mathbf{n}_i$, $i = 1,2,3$, with \mathbf{n}_i being the dextral unit vectors of the frame \mathbf{x} .

The cross sectional area can be slowly varying function of s and will be denoted by $A(s)$. As the rod deforms the curve C acquires new configuration c that changes with time. The arc length along c is denoted by ξ that depends upon s and t , i.e. $\xi = \xi(s,t)$. The position of a point s on c at an arbitrary time is $\mathbf{x}(s,t)$ so that

$$\mathbf{x}(s,t_0) = \mathbf{X}(s), \quad \xi(s,t_0) = s \quad (1a,b)$$

where t_0 is a reference time at which the rod is in the unstrained state. Also

$$\frac{\partial \xi}{\partial s} = \xi' = (\mathbf{x}'_i \mathbf{x}'_i)^{\frac{1}{2}} \quad (2)$$

where prime denotes differentiation with respect to time and summation over a repeated index is implied. The strain e is defined by

$$e = \xi' - 1 \quad (3)$$

where $e > 0$ denotes extension and $e < 0$ contraction. The strict positivity of ξ' implies that $-1 \leq e < \infty$.

Attached to any point s of c a Cartesian coordinate frame y will be assumed and will be referred to as the body reference frame. The coordinate axes of the body frame are y_1, y_2, y_3 with the y_3 axis pointing in the direction of increasing s and y_1 and y_2 being the principal axes of the cross section. The dextral unit vectors of the y frame will be denoted by $e_i, i = 1, 2, 3$, with their orientations at the reference time t_0 being E_i .

Denoting by $l_{ij}(s, t)$ and $L_{ij}(s)$ the elements of the matrices of direction cosines of the dextral sets e_i and E_i one has

$$n_i = l_{ij} e_j = L_{ij} E_j \quad (4)$$

and

$$e_i = l_{ij} n_j, \quad E_i = L_{ij} n_j \quad (5a, b)$$

The angles φ_i represent rotations between the corresponding pairs of E_i and e_i when these directions are assumed to issue from a common origin, Fig. 1. These angles are determined through

$$\cos \varphi_i = e_i \cdot E_i = l_{ji} L_{ji}, \quad i=1, 2, 3 \text{ (sum only on } j) \quad (6)$$

The direction cosines l_{ij} are characterized non-uniquely by three orientations angles $\theta_1(s, t)$, $\theta_2(s, t)$ and $\theta_3(s, t)$. These angles can be selected in a variety of ways and represent three finite rotations about unit vectors e_i or n_i . If these rotations are properly selected the fixed orientation n_i may be brought to any arbitrary body orientation e_i . Kane et al. (1983) list at least 24 possibilities for the order of rotations of the angles $\theta_1, \theta_2, \theta_3$ about the body set of unit vectors e_i or the fixed set of unit vectors n_i .

Regardless of the particular choice of orientation angles the angular velocity $\omega = \omega_i e_i$ of a cross sectional element $Ad s$, Fig. 1 of the rod is determined uniquely from

$$\omega_i = \eta_{igh} \dot{l}_{ig} l_{ih}, \quad (\dot{} = \frac{\partial}{\partial t}) \quad (7)$$

where $\eta_{ijk} = \epsilon_{ijk} (\epsilon_{ijk} + 1)/2$, (no sum on i, j, k) and ϵ_{ijk} is the alternator tensor with the non-zero components $\epsilon_{123} = \epsilon_{231} = \epsilon_{321} = +1$ and $\epsilon_{132} = \epsilon_{321} = \epsilon_{213} = -1$.

The curvature vector $K = K_i e_i$ of the rod can be defined in a similar way with K_3 representing twist and K_1 and K_2 representing bending curvatures about the principal directions of the cross section. Using the dynamical analogy of E.I. Routh, Love (1944) has noted that if the frame y were to move with unit speed along the curve c such that at any point ξ of c it has the orientation of the y frame at that point then the angular velocities ω_1 and ω_2 will be the principal curvatures K_1 and K_2 of the rod. Also the angular velocity ω_3

will be the twist curvature K_3 of the rod. Thus the formulas that define the angular velocities from direction cosines can be used to determine curvatures, provided time differentiation is replaced by differential with respect to ξ . Therefore one has

$$K_i = \eta_{igh} \frac{\partial l_{ig}}{\partial \xi} l_{ih} = \frac{1}{1+e} \eta_{igh} l'_{ig} l_{ih} \equiv \frac{k_i}{1+e}, \quad (8)$$

where differentiation with respect to ξ has been replaced with differentiation with respect to s and curvature parameters $k_i = (1+e)K_i$ is also introduced. For future use one may note the formulas for derivatives of direction cosines l_{ij} and the unit vectors e_i .

$$l_{ij} = (1+e) \epsilon_{ghj} l_{ig} k_h \quad \dot{l}_{ij} = \epsilon_{ghj} l_{ig} w_h \quad (9a,b)$$

$$\dot{e}_i = \epsilon_{kji} k_j e_k \quad \dot{e}_i = \epsilon_{kji} w_j e_k \quad (10a,b)$$

if K_i be the curvatures of the rod in the unstrained state ($e = 0$) then from (8)

$$K_i = \eta_{igh} L'_{ig} L_{ih} \quad (11)$$

Since \dot{e}_3 and $\partial x / \partial \xi$ are both unit tangents to the central line one has

$$\dot{x}_i = (1+e) l_{i3} \quad (12)$$

We assume that the center of mass of the cross sections coincide with the centroids. The linear and the central angular momentum per unit length are then given by

$$p = \rho A x_i n_i \quad (13)$$

and

$$H = \rho I \omega = \rho (I_{11} \omega_1 e_1 + I_{22} \omega_2 e_2 + I_{33} \omega_3 e_3) \quad (14)$$

where ρ is the mass density per unit unstrained length and I is the diagonal moment of inertia tensor with components

$$I_{11} = \int_A y_2^2 dA, \quad I_{22} = \int_A y_1^2 dA, \quad I_{33} = I_{11} + I_{22} \quad (15)$$

Equations of Motion

Referring to the body set of axes e_i one can define the vector F of the resultant shear stresses F_1 and F_2 and the axial stress resultant F_3 . Similarly, one may define the couple stress vector M consisting of the bending moments M_1 and M_2 and the torque M_3 . Explicitly we have

$$F = F_i e_i \quad \text{and} \quad M = M_i e_i \quad (16)$$

The well known dynamic equilibrium of the rod can be expressed by the equations of the balance of linear momentum

$$\mathbf{F}' + \mathbf{f} = \dot{\mathbf{p}} \quad (17)$$

and of the balance of angular momentum

$$\mathbf{M}' + \mathbf{x}' \times \mathbf{F} + \mathbf{m} = \dot{\mathbf{H}} \quad (18)$$

wherein \mathbf{f} and \mathbf{m} represent distributed force and moment acting on the rod. The scalar components of these equations can be referred to the body set of axes. For this purpose one needs to express all vector quantities in terms of unit vectors \mathbf{e}_i and use (9)–(10). Then (17) becomes

$$F'_1 + k_2 F_3 - k_3 F_2 + f_1^y = \rho A \bar{x}_j l_{j1} \quad (19a)$$

$$F'_2 + k_3 F_1 - k_1 F_3 + f_2^y = \rho A \bar{x}_j l_{j1} \quad (19b)$$

$$F'_3 + k_1 F_2 - k_2 F_1 + f_3^y = \rho A \bar{x}_j l_{j3} \quad (19c)$$

while (18) assumes the form

$$M'_1 + k_2 M_3 - k_3 M_2 - (1+e)F_2 + m_1^y = \rho I_1 (\omega_1 + \omega_2 \omega_3) \quad (20a)$$

$$M'_2 + k_3 M_1 - k_1 M_3 - (1+e)F_1 + m_2^y = \rho I_2 (\omega_2 + \omega_1 \omega_3) \quad (20b)$$

$$M'_3 + k_1 M_2 - k_2 M_1 + m_3^y = \rho [J \dot{\omega}_3 + (I_2 - I_1) \omega_1 \omega_2] \quad (20c)$$

where $I_1 = I_{11}$, $I_2 = I_{22}$ and $I = J_{33} = I_1 + I_2$. The superscript y on the components of \mathbf{f} and \mathbf{m} denote the components of these vectors in the body reference frame.

To express the equations of motion in the inertial frame we introduce the components of the stress resultants in that frame. Thus

$$\mathbf{F}_i^x = l_{ij} F_j \quad \mathbf{M}_i^x = l_{ij} M_j \quad (21a,b)$$

Then (19) becomes

$$F_i^x + f_i^x = \rho A \ddot{x}_i \quad (22)$$

and (20) assumes the form

$$M_1^{x'} - (1+e)F_2 + m_1^x = \rho (I_{rj} l_{ij} \dot{\omega}_r + \epsilon_{grj} I_{sj} l_{ig} \omega_s \omega_r) \quad (23a)$$

$$M_2^{x'} - (1+e)F_1 + m_2^x = \rho (I_{rj} l_{2j} \dot{\omega}_r + \epsilon_{grj} I_{sj} l_{2g} \omega_s \omega_r) \quad (23b)$$

$$M_3^{x'} + m_3^x = \rho(I_{Tj} l_{3j} \dot{\omega}_r + \epsilon_{grj} I_{sj} l_{3g} \omega_s \omega_r) \quad (23c)$$

Either of the set of equations (19)–(20) or (22)–(23) can be considered as the governing differential equations of motion. These equations will have to be supplemented with constitutive relations that define resultant axial stress F_3 and the resultant bending and twisting couples M_1, M_2, M_3 in terms of the axial strain e and curvatures k_1, k_2, k_3 . In the next section we consider the derivation of these constitutive relations.

Constitutive Relations

We assume that the motion of the elastic rod is equivalent to the stationarity of the Hamiltonian H which is defined by

$$\begin{aligned} H[l_{ij}(\varphi), x_i, e] = & \int_{t_1}^{t_2} \int_{s_1}^{s_2} \mathcal{L} ds dt + \\ & \int_{t_1}^{t_2} \left[\bar{F}_i \bar{X}_i + \bar{M}_i \bar{\varphi}_i \right]_{s_1}^{s_2} dt - \int_{s_1}^{s_2} \left[\rho A \bar{v}_i x_i \right]_{t_1}^{t_2} ds - \\ & \int_{s_1}^{s_2} \rho \left[I_1 \bar{\omega}_1 \varphi_1 + I_2 \bar{\omega}_2 \varphi_2 + J \bar{\omega}_3 \varphi_3 \right]_{t_1}^{t_2} ds \end{aligned} \quad (24)$$

where \mathcal{L} is the action density function

$$\begin{aligned} \mathcal{L} = & \frac{1}{2} \rho A \dot{x}_i \dot{x}_i + \frac{1}{2} \rho (I_1 \dot{\omega}_1^2 + I_2 \dot{\omega}_2^2 + J \dot{\omega}_3^2) - w(e, k_i) + \\ & \lambda_i [l_{i3} (1+e) - x_i'] + f_i^x x_i + m_i^y \varphi_i \end{aligned} \quad (25)$$

and \bar{F}_i^1, \bar{M}_i^1 and \bar{F}_i^2, \bar{M}_i^2 are the applied forces and moments at ends s_1 and s_2 respectively.

Also $\bar{v}_i^{1,2}$ and $\bar{\omega}_i^{1,2}$ are the initial and final linear and angular velocities. The strain energy function w depends upon the kinematical variables e and k_i . The precise nature of this dependence is the constitutive relations that we seek and is a consequence of the stationarity of H . The functions λ_i are the Lagrange multipliers that allow the constraints (12) to be incorporated within the Hamiltonian. As a result x_i and l_{ij} can be regarded as independent variables. Additionally the constraint (12) implies the definition (2)–(3) for the strain e and hence in (24) e can also be viewed as an independent variable. To see this we need to note that if each side of (12) is multiplied by itself we obtain $x_i' x_i' =$

$(1+e)^2 l_{i3} l_{i3} = (1+e)^2$ which is restatement of (2)–(3). The terms $f_i^x x_i$ and $m_i^y \varphi_i$ in (25) represent the density of the potential of the applied forces and moments on the rod. As stated in (5) the angles φ_i are the rotations from E_i to e_i .

With these preliminaries we note that the Euler equation corresponding variations δx_i is simply $\lambda_i' + f_i^x = \rho A \bar{x}_i$ which when compared with (22) reveals that $\lambda_i = F_i^x$. Next considering the variations with respect to e we obtain

$$\frac{\partial W}{\partial e} = F_i^x l_{i3} = F_3 \quad (26)$$

which is the constitutive relationship determining the axial force F_3 as the derivative of strain energy with respect to axial strain e .

We now turn to the Euler equation corresponding to the variation $\delta \varphi_1$. For this purpose we note that l_{ij} , ω_i and k_i depend on φ_1 . For orientation angles of the cross section we select the sequence of body rotations first $\theta_2 e_2$, second $\theta_3 e_3$ and the third $\theta_1 e_1$ with $\theta_1 \equiv \varphi_1$. In this sequence the last rotation is through φ_1 with respect to which variation is sought. The matrix l of the direction cosines is given by

$$l = B(\theta_2)C(\theta_3)A(\theta_1) = \begin{bmatrix} C_2 C_3 & -C_1 C_2 S_3 + S_1 S_2 & S_1 C_2 S_3 + C_1 S_2 \\ S_3 & C_1 C_3 & -S_1 C_3 \\ -S_2 C_3 & C_1 S_2 S_3 + S_1 C_2 & -S_1 S_2 S_3 + C_1 C_2 \end{bmatrix} \quad (27)$$

where

$$C_i = \cos \theta_i, \quad S_i = \sin \theta_i \quad (28)$$

$$A(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos \theta & -\sin \theta \\ 0 & \sin \theta & \cos \theta \end{bmatrix} \quad B(\theta) = \begin{bmatrix} \cos \theta & 0 & \sin \theta \\ 0 & 1 & 0 \\ -\sin \theta & 0 & \cos \theta \end{bmatrix}$$

$$C(\theta) = \begin{bmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (29)$$

Subsequently, we find from (7) and (8)

$$\omega_1 = \dot{\theta}_2 S_3 + \dot{\theta}_1, \quad \omega_2 = \dot{\theta}_2 C_1 C_3 + \dot{\theta}_3 S_1, \quad \omega_3 = -\dot{\theta}_2 S_1 C_3 + \dot{\theta}_3 C_1 \quad (30)$$

$$k_1 = \dot{\theta}_2' S_3 + \dot{\theta}_1', \quad k_2 = \dot{\theta}_2' C_1 C_3 + \dot{\theta}_3' S_1, \quad k_3 = -\dot{\theta}_2' S_1 C_3 + \dot{\theta}_3' C_1 \quad (31)$$

From (24) we have

$$\left\{ -\frac{\partial W}{\partial k_1} \frac{\partial k_1}{\partial \varphi_1} + \frac{\partial}{\partial s} \left(\frac{\partial W}{\partial k_1} \frac{\partial k_1}{\partial \varphi_1} \right) + (1+e) F_1^x \left[\frac{\partial l_{13}}{\partial \varphi_1} - \frac{\partial}{\partial s} \left(\frac{\partial l_{13}}{\partial \varphi_1} \right) \right] \right\} + m_1^y = \frac{\partial}{\partial t} \left[\rho I_1 \omega_1 \frac{\partial \omega_1}{\partial \varphi_1} \right] - \rho I_2 \omega_2 \frac{\partial \omega_2}{\partial \varphi_1} - \rho J \omega_3 \frac{\partial \omega_3}{\partial \varphi_1} \quad (32)$$

Noting that $\theta_1 \equiv \varphi_1$ we have from (30) $\partial \omega_1 / \partial \varphi_1 = 1$, $\partial \omega_2 / \partial \varphi_1 = \omega_3$, $\partial \omega_3 / \partial \varphi_1 = -\omega_2$. Also from (31) $\partial k_1 / \partial \varphi_1 = 0$, $\partial k_2 / \partial \varphi_1 = k_3$, $\partial k_3 / \partial \varphi_1 = -k_2$ and from (27) $\partial l_{13} / \partial \varphi_1 = -l_{12}$. Using these results and the inverse of (21a), (32) becomes

$$\left(\frac{\partial W}{\partial k_1} \right)' + k_2 \frac{\partial W}{\partial k_3} - k_3 \frac{\partial W}{\partial k_2} - (1+e) F_2 + m_1^y = \rho I_1 (\dot{\omega}_1 + \omega_2 \omega_3) \quad (33)$$

In exactly the same manner one can proceed to determine the Euler equation for a variation $\delta \varphi_2$. Now the consecutive sequence of body rotations $\theta_3 e_3$, $\theta_1 e_1$ and $\theta_2 e_2$ is selected with $\theta_2 \equiv \varphi_2$. Without going into details one obtains

$$\left(\frac{\partial W}{\partial k_2} \right)' + k_3 \frac{\partial W}{\partial k_1} - k_1 \frac{\partial W}{\partial k_3} - (1+e) F_1 + m_2^y = \rho I_2 (\dot{\omega}_2 + \omega_1 \omega_3) \quad (34)$$

For variation of φ_3 we adopt the consecutive sequence of body rotations $\theta_1 e_1$, $\theta_2 e_2$ with $\theta_3 \equiv \varphi_3$. The matrix of direction cosines is

$$l = A(\theta_1) B(\theta_2) C(\theta_3) = \begin{bmatrix} C_2 C_3 & -C_2 S_3 & S_2 \\ S_1 S_2 C_3 + C_1 S_3 & -S_1 S_2 S_3 + C_1 C_3 & -S_1 C_2 \\ -C_1 S_2 C_3 + S_1 S_3 & C_1 S_2 S_3 + S_1 C_3 & C_1 C_2 \end{bmatrix} \quad (35)$$

with angular velocities of the cross section and the curvatures given by

$$\omega_1 = \dot{\theta}_1 C_2 C_3 + \dot{\theta}_2 S_3, \quad \omega_2 = \dot{\theta}_2 C_3 - \dot{\theta}_1 C_2 S_3, \quad \omega_3 = \dot{\theta}_3 + \dot{\theta}_1 S_2 \quad (36)$$

$$k_1 = \theta_1' C_2 C_3 + \theta_2' S_3, \quad k_2 = \theta_2' C_3 - \theta_1' C_2 S_3, \quad k_3 = \theta_3' + \theta_1' S_2 \quad (37)$$

For this case l_{13} does not depend on θ_3 and hence the Euler variational equation assumes the form

$$\left(\frac{\partial W}{\partial k_3} \right)' + k_1 \frac{\partial W}{\partial k_2} - k_2 \frac{\partial W}{\partial k_1} + m_3^y = \rho [J \dot{\omega}_3 + (I_2 - I_1) \omega_1 \omega_2] \quad (38)$$

The specified boundary conditions at $s = s_1, s_2$ must be consistent with

$$[(\bar{F}_1 - F_1^x) \delta x_1 + (\bar{M}_1 - M_1) \delta \varphi_1]_{s_1}^{s_2} = 0 \quad (39)$$

for arbitrary and independent variations δx_1 and $\delta \varphi_1$. Similar restrictions are imposed on initial and final data, i.e.

$$\left\{ \rho A(\dot{x}_i - \bar{v}_i) \delta x_i + \rho [I_1(\omega_1 - \bar{\omega}_1) \delta \varphi_1 + I_2(\omega_2 - \bar{\omega}_2) \delta \varphi_2 + J(\omega_3 - \bar{\omega}_3) \delta \varphi_3] \right\}_{t_1}^{t_2} = 0 \quad (40)$$

Comparison of equations (33), (34) and (38) with equations (20a,b,c) respectively, establishes the constitutive relations

$$M_i = \frac{\partial W}{\partial k_i} \quad i = 1, 2, 3 \quad (41)$$

A STRAIN ENERGY FUNCTION

In order to gain an insight into the nature of the strain energy function we consider the strain of the lines and angles in the cross section of the rod. For this purpose we invoke the Kirchhoff hypothesis which assumes that the plane cross sections of rod that are normal to the axial direction in the unstrained state remain normal to the strained axial direction during deformation. Therefore the position vector to a material point in the cross-section before and after deformation can be given by

$$\mathbf{R} = \mathbf{X}(s) + y_1 \mathbf{E}_1(s) + y_2 \mathbf{E}_2(s) \quad (42)$$

and

$$\mathbf{r} = \mathbf{x}(s) + \alpha(s)[y_1 \mathbf{e}_1(s) + y_2 \mathbf{e}_2(s)] \quad (43)$$

respectively. The parameter $\alpha(s)$ is to be fixed by enforcing traction-free boundary conditions on the lateral surface of the rod.

Using the concept of extensional strains for stretching of line elements and distortion of angles between perpendicular lines as shear strains (Wempner, 1991), we define components of strain by

$$\epsilon_{ij} = \frac{1}{2}(\mathbf{g}_i \cdot \mathbf{g}_j - \mathbf{G}_i \cdot \mathbf{G}_j) \quad (44)$$

where

$$\begin{aligned} \mathbf{g}_1 &= \frac{\partial \mathbf{r}}{\partial y_1} = \alpha \mathbf{e}_1, \quad \mathbf{g}_2 = \frac{\partial \mathbf{r}}{\partial y_2} = \alpha \mathbf{e}_2, \\ \mathbf{g}_3 &= \frac{\partial \mathbf{r}}{\partial y_3} = \alpha' y_1 \mathbf{e}_1 + \alpha' y_2 \mathbf{e}_2 + \alpha y_1 \mathbf{e}_1' + \alpha y_2 \mathbf{e}_2' + (1 + \epsilon) \mathbf{e}_3 \end{aligned} \quad (45)$$

$$\mathbf{G}_1 = \frac{\partial \mathbf{R}}{\partial y_1} = \mathbf{E}_1, \quad \mathbf{G}_2 = \frac{\partial \mathbf{R}}{\partial y_2} = \mathbf{E}_2,$$

$$\mathbf{G}_3 = \frac{\partial \mathbf{R}}{\partial y_3} = y_1 \mathbf{E}_1' + y_2 \mathbf{E}_2' + \mathbf{E}_3 \quad (46)$$

Using (8) we can establish

$$\mathbf{e}_i' \cdot \mathbf{e}_j = \epsilon_{ij} \mathbf{k}_n \quad (47)$$

$$\mathbf{E}_i' \cdot \mathbf{E}_j = \epsilon_{ij} K_n \quad (48)$$

where K_i is the curvatures and twist in the unstrained state. Therefore (9) yields as the strain components

$$\begin{aligned} \epsilon_{11} &= \epsilon_{22} = \frac{1}{2}(\alpha^2 - 1), \quad \epsilon_{12} = 0 \\ \epsilon_{13} &= \frac{1}{2}[\alpha(\alpha' y_1 - \alpha y_2 k_3) + y_2 K_3] \\ \epsilon_{23} &= \frac{1}{2}[\alpha(\alpha' y_2 + \alpha y_1 k_3) - y_1 K_3] \\ \epsilon_{33} &= e + \frac{1}{2}e^2 - y_1[(1+e)\alpha k_2 - K_2] + y_2[(1+e)\alpha k_1 - K_1] - y_1 y_2(\alpha^2 k_1 k_2 - K_1 K_2) + \\ &\quad \frac{1}{2}y_1^2[\alpha'^2 + \alpha^2(k_2^2 + k_3^2) - K_2^2 - K_3^2] \\ &\quad + \frac{1}{2}y_2^2[\alpha'^2 + \alpha^2(k_1^2 + k_3^2) - K_1^2 - K_3^2] \end{aligned} \quad (49)$$

For a linear isotropic elastic material the non-zero stress components per unit strained area are

$$\begin{aligned} \sigma^{11} &= (\lambda + G)(\alpha^2 - 1) + \lambda \epsilon^{33}, \quad \sigma^{12} = 0 \\ \sigma^{22} &= (\lambda + G)(\alpha^2 - 1) + \lambda \epsilon^{33}, \quad \sigma^{23} = 2G\epsilon_{23} \\ \sigma^{33} &= \lambda(\alpha^2 - 1) + (\lambda + 2G)\epsilon^{33}, \quad \sigma^{31} = 2G\epsilon_{31} \end{aligned} \quad (50)$$

where λ and G are Lamé's constant and the shear modulus, respectively. The traction per unit undeformed area is then given by

$$\mathbf{t}^3 = \sigma^{3i} \mathbf{g}_i \quad (51)$$

One can define the axial stress resultant F_3 by

$$\begin{aligned} F_3 &= \int_A \mathbf{t}^3 \cdot \mathbf{e}_3 \, dA = \\ &= A(1+e) \left[(\lambda + 2G)(e + \frac{1}{2}e^2) + \lambda(\alpha^2 - 1) \right] \\ &\quad + (\lambda + 2G) \left[I_1[(1+e)\alpha k_1 - K_1]\alpha k_1 + I_2[(1+e)\alpha k_2 - K_2]\alpha k_2 \right. \\ &\quad \left. + \frac{1}{2}I_1(1+e)(\alpha'^2 + \alpha^2 k_1^2 - K_1^2 + \alpha^2 k_3^2 - K_3^2) \right. \\ &\quad \left. + \frac{1}{2}I_2(1+e)(\alpha'^2 + \alpha^2 k_2^2 - K_2^2 + \alpha^2 k_3^2 - K_3^2) \right] \end{aligned} \quad (52)$$

Similarly we have

$$M_1 = \int_A \alpha y_2 t^3 \cdot e_3 dA = \alpha(\lambda+2G)I_1[(1+e)\alpha k_1 - K_1](1+e) \\ + I_1 \alpha^2 k_1 [\lambda(\alpha^2 - 1) + (\lambda + 2G)(e + \frac{1}{2}e^2)] \quad (53)$$

$$M_2 = -\int_A \alpha y_1 t^3 \cdot e_3 dA = \alpha(\lambda+2G)I_2[(1+e)\alpha k_2 - K_2](1+e) \\ + I_2 \alpha^2 k_2 [\lambda(\alpha^2 - 1) + (\lambda + 2G)(e + \frac{1}{2}e^2)] \quad (54)$$

$$M_3 = -\int_A \alpha(y_1 t^3 \cdot e_3 - y_2 t^3 \cdot e_1) dA = JG\alpha^2(\alpha^2 k_3 - K_3) \\ + J\alpha^2 k_3 [\lambda(\alpha^2 - 1) + (\lambda + 2G)(e + \frac{1}{2}e^2)] \quad (55)$$

One may note that the integrability conditions

$$\frac{\partial F_3}{\partial k_1} = \frac{\partial M_1}{\partial e}, \quad \frac{\partial F_3}{\partial k_2} = \frac{\partial M_2}{\partial e}, \quad \frac{\partial F_3}{\partial k_3} = \frac{\partial M_3}{\partial e}, \\ \frac{\partial M_1}{\partial k_2} = \frac{\partial M_2}{\partial k_1}, \quad \frac{\partial M_1}{\partial k_3} = \frac{\partial M_2}{\partial k_1}, \quad \frac{\partial M_2}{\partial k_3} = \frac{\partial M_3}{\partial k_2} \quad (56)$$

are satisfied. Hence existence of a strain energy function is assured and by integration we have

$$W = A \frac{\lambda+2G}{2} e^2 (1+e+\frac{e^2}{4}) - A\lambda(1-\alpha^2)(e+\frac{1}{2}e^2) \\ + \lambda I_1 \alpha^2 (\alpha^2 - 1) \frac{k_1^2}{2} + (\lambda+2G)I_1 \alpha k_1 \left[\frac{1+2e^2+4e}{4} \alpha k_1 - (1+e)K_1 \right] \\ + \lambda I_2 \alpha^2 (\alpha^2 - 1) \frac{k_2^2}{2} + (\lambda+2G)I_2 \alpha k_2 \left[\frac{1+2e^2+4e}{4} \alpha k_2 - (1+e)K_2 \right] \\ + \frac{\lambda+2G}{2} (1+e)^2 \left[I_1 (\alpha'^2 + \alpha^2 k_1^2 - K_1^2) + I_2 (\alpha'^2 + \alpha^2 k_2^2 - K_2^2) \right] \\ + \lambda J \alpha^2 (\alpha^2 - 1) \frac{k_3^2}{2} + \frac{\lambda+2G}{2} J (e+\frac{1}{2}e^2) (\alpha^2 k_3^2 - K_3^2) + \frac{1}{2} G J \alpha (\alpha k_3 - K_3)^2 \quad (57)$$

For an initially straight rod K_i should be set equal to zero. The above form of W reflects material isotropy, i.e. $W(e, k_1, k_2, k_3) = W(e, k_2, k_1, k_3)$, provided that $I_1 = I_2$.

We note that positive curvatures imply positive bending moments and conversely negative curvatures imply negative bending moments provided that $e > (-1 + 1/\sqrt{3})$ for $\alpha = 1$. This shows that equations (53) have a limited range of validity if the sense correspondence between moments and curvatures is to be retained. We also note the second order coupling between the squares of the curvatures and the axial strain e in (52), which implies that axial force can be generated by bending or twist only.

The parameter $\alpha(s)$ depends on the boundary conditions applied at the lateral surface of the rod. If the lateral surface is fixed, then $\alpha(s) = 1$. For zero tractions on the lateral surface, a condition appropriate for thin flexible rods is adopted according to which the average of σ^{11} and σ^{22} over the cross-section should vanish, i.e.

$$\int_A \sigma^{11} dA = \int_A \sigma^{22} dA = 0 \quad (58)$$

The above condition reduces to

$$\begin{aligned} H(\alpha, k_i, e) &\equiv \alpha' J + \alpha^2 (I_1 k_1^2 + I_2 k_2^2 + J k_3^2 + \frac{A}{\nu}) \\ &- (I_1 K_1^2 + I_2 K_2^2 + J K_3^2) + 2Ae(1 + \frac{1}{2}e) - \frac{A}{\nu} = 0 \end{aligned} \quad (59)$$

This equation should be interpreted as a differential equation for $\alpha(s)$ when k_i and e are known. To achieve this (52) is solved in an iterative procedure in which at every step k_i and e are known. We begin by writing (52) as $\lim_{n \rightarrow \infty} H(\alpha_n, k_i^n, e_n) = 0$ with $n = 0, 1, 2, 3, \dots$

... For the first iteration ($n=0$), $e_0 = 0$, $k_i^0 = K_i$, $\alpha_1 = 1$, and the six equations (15) – (16) after using (22), (33) and (57) contain only six unknown quantities F_1 , F_2 , e_1 and k_i^1 when the externally applied forces and moments are prescribed. Solution of this set of equations enables one to use k_i^1 in the curvature-orientation angle relations such as (31) to determine the latter i.e. θ_i . Now l_{ij} (θ_k) are known and one proceeds to determine φ_i from (21) and x_i from (9) using the appropriate boundary conditions. The solution for the first iteration is complete. One enters (52) with e_1 and k_i^1 and computes α_2 and the iteration proceeds.

Acknowledgement

This work was supported by the Army Research Office grant No. DAAL – 03 – 92 G – 0123 to Rensselaer Polytechnic Institute.

REFERENCES

- Antman, S.S., "Ordinary Differential Equations of Non-Linear Elasticity I: Foundations of the Theories of Non-Linearly Elastic rods and Shells," A.R.M.A. 61 (1976), 307-351.
- Antman, S.S., "The Theory of Rods", Handbuch der Physik, Vol. VIa/2, Springer-Verlag, Berlin, 1972.
- Cosserat, E. and F., "The Des Corps Deformables, A. Hermann et Fils, Paris, 1909.
- Ericksen, J.L., "Simpler Static Problems in Nonlinear Theories of Rods," Int. J. Solids Structures 6 (1970) 371-377.
- Green, A.E. and L. Laws, "A General Theory of Rods", Proc. Royal Soc. Lond. A293 (1966) 145-155.
- Green, A.E., P.M. Naghdi and M.L. Wenner, "On the Theory of Rods. I Derivations from the Three-Dimensional Equations and II Developments by Direct Approach," Proc. Royal Society London A337 (1974) 451-507.
- Kane, T.R., Likins, P.W. and Levinson, D.A., "Spacecraft Dynamics," McGraw-Hill, New York, 1983.
- Love, A.E.H., "A Treatise of the Mathematical Theory of Elasticity", Dover, 1944.
- Naghdi, P.M. and M.B. Rubin, "Constrained Theories of Rods", J. Elasticity 14 (1984) 343-361.
- Naghdi, P.M., "Finite Deformation of Elastic Rods and Shells," Proc. IUTAM Symp. - Finite Elasticity, Bethlehem, PA, 1980 (Edited by D.E. Carlson and R.T. Shield), 47-103, Martinus Nijhoff, The Hague, 1982.
- Poynting, J.H., "On Pressure Perpendicular to the Shear Planes in Finite Pure Shears, and on the Lengthening of Loaded Wires When Twisted," Proc. Roy. Soc. London A82 (1909) 546-559.
- Rivlin, R.S. and D.W. Saunders, "Large Elastic Deformations of Isotropic Materials. VII. Experiments on the Deformation of Rubber," Phil. Trans. Roy. Soc. London, Ser. A, No. 865, 243 (1951) 251-288.
- Tadibakhsh, I., "The Variational Theory of the Plane Motion of the Extensible Elastica", Int. J. Engng. Sci. 4 (1966) 433-450.
- Wempner, G., "Mechanics of Deformable Solids," Wempner, 1991.
- Whitman, A.B. and C.N. DeSilva, "Dynamics and Stability of Elastic Cosserat Curves", Int. J. Solids Structures 6 (1970) 411-422.

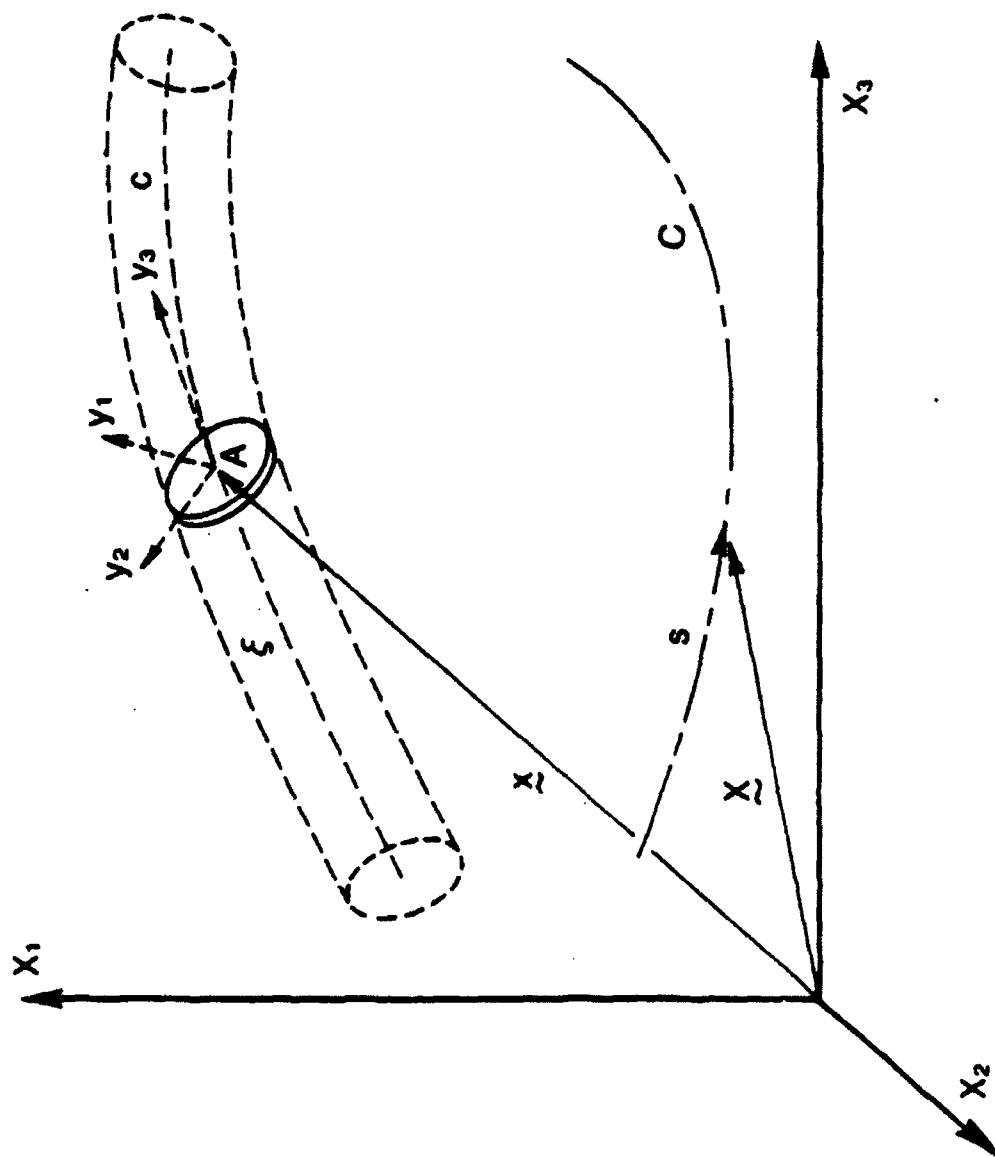


Fig. 1

VISCOHYPERELASTICITY

A. R. Johnson^{*}, C. J. Quigley^{**} and C. E. Freese^{**}

Army Research Laboratory
Watertown, MA 02172-0001

ABSTRACT

A recently developed^{1,2,3,4} internal solid theory for rubber viscoelasticity is reviewed in the context of finite hyperelasticity. Viscohyperelasticity, in which time dependent reference geometries are used with internal solids having hyperelastic energy functions, is developed. The material constants for the internal solids are determined by tensile, shear and biaxial tests. Computational aspects of the theory and its inclusion into finite element analysis are discussed here. Previous finite element models^{1,2,3,4} are improved with a predictor-corrector scheme. Time dependent pressure loads are applied to the interior of a thick walled rubber cylinder to demonstrate the use of viscohyperelasticity.

INTRODUCTION

The development of useful computational models for rubber vulcanizates requires interactions between experimentalists and analysts. These interactions often involve modeling the performance of rubber subjected to dynamic strains. Many studies have been done to analyze the case of small dynamic strains about a large strain deformed elastic state. Although the information obtained from such studies has proven valuable for the design process, it does not preclude the need for a capability to measure and predict large strain dynamic data. Viscoelastic models valid for large strains have been under development for about forty years. Typical weaknesses of these models are a lack of standard laboratory tests to determine their parameters and their inability to predict a wide range of large strain dynamic test data.

* Vehicle Structures Division

** Mechanics Division

History integral formulations for modeling the viscoelastic response of elastomers have been used by Bernstein, Kearsley and Zapas⁵ and McGuirt and Lianis^{6,7}. Their efforts consisted of applying the theory of Green and Rivlin⁸ to the viscoelastic deformations of both low and high percentage crosslinked elastomers. In both cases they demonstrated that only a few parameters are needed to model most materials of interest. However, the history integral method has not proved attractive for finite element algorithms due to its large storage and computational requirements. The fact that only a few parameters were needed to model viscoelastic effects in the rubberlike materials they tested implies that alternative models may also be useful. Finite element viscoelasticity algorithms using Kelvin elements (valid for small strains) were developed by Zienkiewicz, Watson and King⁹, and Carpenter¹⁰. These algorithms avoided the use of the history integral method, are computationally attractive and have given acceptable accuracy for small strains.

In this paper we review viscohyperelasticity^{1,2,3,4} and present a finite element analysis of a thick viscoelastic rubber cylinder subjected to a cyclic internal pressure loading. We also restrict this paper to the essentials of the computational algorithm.

VISCOHYPERELASTICITY

Background information on hyperelasticity, determining principal stretches, strain invariants, etc. is available in many books (for example, see Treloar¹¹, Ogden¹², or Green and Zerna¹³). The first new concept to describe in viscohyperelasticity is that of a changing reference shape. This can be seen by considering the step-strain relaxation of a two network system consisting of chemically crosslinked molecules and entangled molecules as shown in Figure 1. The sides of the rectangles represent the chemically crosslinked molecules which are not damaged during deformation. The diagonal represents a molecule which is rigidly bonded at its ends to the chemically crosslinked network and the loop on the diagonal represents an entanglement which can slip when stressed. During the step-strain relaxation the stress due to the deformation of the chemically crosslinked system (square \rightarrow rectangle) remains constant in time. It is determined from a hyperelastic energy function. The stress, σ in Figure 1, due to the deformation of the

entangled molecule (or a molecule which breaks and reforms continuously in time) decreases in time since the reference shape continuously changes until it looks like the deformed shape (see the bottom of Figure 1). The second concept is related to Green and Tobolsky's¹⁴ work, it states that the force driving the changes in reference shape is a function of the difference (in the sense of a mapping function) between its current shape and the shape of the chemically crosslinked network to which it is attached. The overview of viscohyperelasticity follows.

Consider a collection of N solids, with reference shapes given by coordinates (X_{s1}, X_{s2}, X_{s3}) , where $s = 1, 2, \dots, N$, and each with the same deformed shape given by coordinates (x_1, x_2, x_3) . Let one-to-one differentiable mappings between the reference shapes and the deformed shape be indicated by

$$D_s = D_s(X_s, x) \quad (1)$$

where $X_s = (X_{s1}, X_{s2}, X_{s3})$, $x = (x_1, x_2, x_3)$, see Figure 2. The mappings of equations (1) are used to determine the principal stretches for each solid, $\lambda_s = (\lambda_{s1}, \lambda_{s2}, \lambda_{s3})$. Let the solids be isotropic and hyperelastic with energy density functions given by

$$W_s = W_s(\lambda_s) \quad (2)$$

Then, for the case of incompressible solids, the principal Cauchy stresses are given by

$$\tau_{si} = \lambda_{si} \frac{\partial W_s}{\partial \lambda_{si}} - p_s \quad i = 1, 2, 3 \quad (3)$$

where p_s = the hydrostatic pressure which is determined either analytically by the stress boundary conditions or numerically by the requirement that the mapping of equation (1) represent an incompressible deformation. In continuum mechanics the study of rubber elasticity (for incompressible deformations) is focused on the mathematics of (1) and (3), and on the physics and mathematics of (2).

Assume the stresses in (3) are viscoelastic. Then, let $s = 1$ represent the long term (relaxed) hyperelastic solid whose reference geometry does not

change in time. That is, τ_{1i} = the relaxed stresses at a specified deformation. Let the remaining solids have time dependent reference geometries. We then have $X_s = X_s(t) = (X_{s1}(t), X_{s2}(t), X_{s3}(t))$ for $s = 2, 3, \dots, N$. The Cauchy stress vector in the time dependent deformed body $x(t) = (x_1(t), x_2(t), x_3(t))$ from all of these solids is

$$\{\tau(t)\} = \{\tau_1(X_1, x(t))\} + \sum_{s=2}^N \{\tau_s(X_s(t), x(t))\} \quad (4)$$

$$\text{where } \{\tau_s\} = \sum_{i=1}^3 \tau_{si} \{e_i\} \quad \text{with } s = 1, 2, 3$$

and $\{e_i\}$ = the Cartesian unit vector for the i 'th principal Cauchy stress.

We will have a viscoelastic model if X_s varies in time such that $\tau(t)$ relaxes. That is, such that $\tau_{si}(t) \rightarrow 0$ for $s = 2, \dots, N$ when $x(t)$ = a constant shape.

The viscohyperelastic model simulates relaxation as follows. At any instant of time let the deformed state $x(t)$ be the relaxed state of $X_s(t)$ for $s = 2, \dots, N$. That is, if $x(t)$ is held constant at x_c then require $X_s(t) \rightarrow x_c$ as $t \rightarrow \infty$. This represents a kinematical relaxation process for which

$$\tau_{si}(X_s(t), x_c) \rightarrow \tau_{si}(x_c, x_c) = 0 \quad \text{as } t \rightarrow \infty \quad (5)$$

Then, from (4), $\{\tau_i(t)\} \rightarrow \{\tau_1(X_1, x_c)\}$ = the long term hyperelastic Cauchy stress vector in the solid as $t \rightarrow \infty$.

At this point we note that there are different ways to enforce the kinematical statement $X_s(t) \rightarrow x_c$. We have selected to drive the kinematics with Cauchy stresses on the shapes $X_s(t)$. These stresses are generated as follows. Let the inverse mapping of (1) be given by

$$d_s = d_s(x, X_s) = D_s^{-1}(X_s, x) \quad \text{for } s = 2, \dots, N \quad (6)$$

Use equations (6) to determine principal stretches for each inverse mapping, $\Lambda_s = (\Lambda_{s1}, \Lambda_{s2}, \Lambda_{s3})$. Note, since we assumed the mapping (1) represents an incompressible deformation then the inverse mapping of (6) also represents an

incompressible deformation. For each mapping (6) define hyperelastic energy density functions

$$\hat{W}_s = \hat{W}_s(\Lambda_s) \quad (7)$$

Then, the principal Cauchy stresses on the geometry X_s with respect to geometry x as a reference are given by

$$\hat{\tau}_{si} = \Lambda_{si} \frac{\partial \hat{W}_s}{\partial \Lambda_{si}} - \hat{p}_s \quad i = 1, 2, 3 \quad (8)$$

where $\hat{p}_s =$ a hydrostatic pressure. The energy function \hat{W}_s is ad hoc in this model. The desire is to determine a \hat{W}_s which will allow large strain viscoelastic stress data to be modeled. Note, Maxwell models are also ad hoc.

The stresses in (8) are of a hyperelastic form and the internal solid model is completed by requiring that $X_s(t) \rightarrow x(t)$ by a relaxation of $\hat{\tau}_{si}$. That is, determine $X(t)$ by integrating the following differential equation which states that the rate of change of $\hat{\tau}_{si}$ per unit time is proportional to $\hat{\tau}_{si}$.

$$-\eta_s \frac{\partial \hat{\tau}_{si}}{\partial t} = \frac{\partial \hat{\tau}_{si}}{\partial X_s} \frac{dX_s}{dt} = \hat{\tau}_{si} \quad (9)$$

Equation (9) is a nonlinear differential equation whose integral determines the time dependent undeformed geometries. To implement the theory we must measure stress relaxation data and use it to determine W_1 , W_s , \hat{W}_s and η_s for $s = 2, \dots, N$. Below, we: (a) present a one dimensional example of the theory in which the form of \hat{W}_s is identical to the form of W_s , (b) review the finite element implementation of viscohyperelastic theory and (c) computationally demonstrate the method by applying a dynamic pressure loading to the interior of a thick walled cylinder.

ONE DIMENSIONAL MODEL

A one dimensional time dependent internal solid (network) is shown in Figure 3. It is assumed that in all states the material is NeoHookean with a shear modulus given by μ . Its time dependent reference length is given by $L(t)$ and its deformed (actual) shape is given by $x(t)$. Following Treloar¹¹,

at any time, t , the force required to hold the material at length $x(t)$ is given by

$$f(t) = \mu \left(\lambda - \frac{1}{\lambda^2} \right) \quad (10)$$

where $\lambda = \frac{x(t)}{L(t)} = \lambda(t)$ = the internal solid's stretch ratio.

Next, we compute a force (Cauchy) on the current reference shape, $L(t)$, by computing a stretch $\Lambda(t) = L(t)/x(t)$. This stretch will go to unity as $L(t) \rightarrow x(t)$. The force is given by

$$g(t) = \mu \left(\Lambda - \frac{1}{\Lambda^2} \right) \quad (11)$$

The model is completed by relaxing $g(t)$ with a time constant η as follows

$$-\eta \frac{\partial g}{\partial t} = -\eta \frac{\partial g}{\partial \Lambda} \frac{d\Lambda}{dt} = g \quad (12)$$

Given $x(t)$ (enforced motion), equation (12) can be numerically integrated and the time dependent viscous force for this internal solid (network) is then given by equation (10). If $f(t)$ is specified then a predictor-corrector algorithm (described below) is useful for obtaining the simultaneous solutions of both (10) and (12).

Figure 4 contains a typical response of the above one dimensional internal solid when it is subjected to a step-strain relaxation test. A more general one dimensional model is easily obtained by superimposing a number of internal solids (different energy functions and relaxation times) with a long term solid (time independent reference shape).

FINITE ELEMENT ANALYSIS

Consider the superposition of solids indicated in Figure 2 with finite element discretizations as shown in Figure 5. Assuming inertial forces are negligibly small, the time dependent potential energy of all the solids in Figure 5 is given by

$$\Pi = \sum_e \Pi_e = \sum_{s=1}^N [W_s(X_s, x) + P_s(X_s, x)] - \phi(x, t) \quad (13)$$

where X_s, x = global vectors of nodal unknowns,
 $W_s(X_s, x)$ = hyperelastic energy function for solid s ,
 $P_s(X_s, x)$ = penalty function to enforce incompressibility
for solid s ,

and $\phi(x, t)$ = work done by applied forces.

The first variation of Π in equation (13) yields

$$\delta \Pi = \left(\sum_{s=1}^N g_s - f \right) \delta x + \sum_{s=2}^N h_s \delta X_s \quad (14)$$

where

$$g_s = \frac{\partial}{\partial x} \left(W_s + P_s \right) = g_s(X_s, x) \quad s = 1, 2, \dots, N$$

$$h_s = \frac{\partial}{\partial X_s} \left(W_s + P_s \right) = g_s(X_s, x) \quad s = 2, 3, \dots, N$$

and

$$f = \frac{\partial \phi}{\partial x} = f(x, t).$$

The variables X_s can not be freely changed. They must change in accordance with their relaxation equations hence, $\delta X_s = 0$. Thus, the equations of motion are given by setting the coefficient of δx in equation (14) to zero. These equations must be solved simultaneously with the relaxation equations. Thus, the solutions are obtained by solving

$$\sum_{s=1}^N g_s(X_s, x) - f(x, t) = 0 \quad (15)$$

and

$$-\eta_s \hat{K}_s(x, X_s) \frac{\partial X_s}{\partial t} = \hat{g}_s \quad s = 2, 3, \dots, N \quad (16)$$

where

$$\hat{g}_s = \frac{\partial}{\partial X_s} (\hat{W}_s + \hat{P}_s)$$

$$\hat{K}_s = \frac{\partial^2}{\partial X_s^2} (\hat{W}_s + \hat{P}_s)$$

$\hat{W}_s(x, X_s)$ = a hyperelastic energy function for the relaxation of X_s ,

and $\hat{P}_s(x, X_s) =$ a penalty function to enforce incompressibility on \hat{W}_s .

The integration of equations (15) and (16) can be accomplished with the following predictor-corrector scheme. Let $x(n)$, $X_s(n)$ and $x(n+1)$, $X_s(n+1)$ be the configurations shown in Figure 5 at times t_n and t_{n+1} . Also let $\Delta t = t_{n+1} - t_n$ and let $x^{(k)}(n+1)$, $X^{(k)}(n+1)$ be estimates of $x(n+1)$, $X(n+1)$ at iteration k . We make a trapezoidal prediction for equation (16) and a Newton - Raphson correction of equation (15) as follows.

A. Initialize $X_s(n+1)$ with an Euler step on the relaxation equation.

$$\Delta X_s(n) = - \frac{\Delta t}{\eta_s} [\hat{K}_s(x(n), X_s(n))]^{-1} \hat{g}_s(x(n), X_s(n)) \quad (17)$$

$$X_s^{(1)}(n+1) = X_s(n) + \Delta X_s(n) \quad (18)$$

B. Initialize $x(n+1)$

$$x^{(1)}(n+1) = x(n) \quad (19)$$

Note: As vectors $x(n)$ are collected equation (19) can be improved with conditional extrapolations using previous states $x(n-1)$, $x(n-2)$, etc.

C. Compute equilibrium at t_{n+1} .

$$G^{(k)}(n+1) = \sum_{s=1}^N g_s(X_s^{(k)}(n+1), x^{(k)}(n+1)) - f^{(k)}(x^{(k)}(n+1), t_{n+1}) \quad (20)$$

D. Check for convergence.

$$|| G^{(k)}(n+1) || < \epsilon \quad (21)$$

E. If converged, go to the next time step, equation (17), if not converged update $x(n+1)$ using a Newton - Raphson step on the equilibrium equation.

$$x^{(k+1)}(n+1) = x^{(k)}(n+1) - [K_G^{(k)}(n+1)]^{-1} G^{(k)}(n+1) \quad (22)$$

where

$$K_G^{(k)}(n+1) = \sum_{s=1}^N K_s(X_s^{(k)}(n+1), x^{(k)}(n+1)) - K_f(x^{(k)}(n+1), t_{n+1})$$

$$K_s = \frac{\partial^2}{\partial x^2} (W_s + P_s) \quad \text{and} \quad K_f = \frac{\partial f}{\partial x}$$

F. Finally, we make a trapezoidal update to $X_s(n+1)$ using $x^{(k+1)}(n+1)$

and the relaxation equation

$$\Delta X_s(n+1) = - \frac{\Delta t}{\eta_s} [\hat{K}_s(x^{(k+1)}(n+1), X_s^{(k)}(n+1))] * \hat{g}_s(x^{(k+1)}(n+1), X_s^{(k)}(n+1)) \quad (23)$$

$$X_s^{(k+1)}(n+1) = X_s(n) + 1/2 (\Delta X_s(n) + \Delta X_s(n+1)) \quad (24)$$

G. Go to equation (20).

The above integration algorithm has the desirable feature of maintaining the volume constraint on all solids as accurately as numerically possible at each time step (given that we have a penalty formulation). We now determine the large time dependent displacement of an inflated thick walled cylinder. The dimensions of the cylinder are shown in Figure 6. It is assumed that the cylinder does not deform in the axial direction. Although this problem can be solved with a one dimensional model we use it to test the axisymmetric finite element code (check for symmetry, etc.). We model the long term hyperelastic response with solid #1 and the Mooney energy function by Oden¹⁵ as modified for a penalty enforcement of the volume constraint by Fried and Johnson¹⁶. Two internal solids were used, one quick and one slow relaxing solid. The energy function for each solid was ($s = \text{solid \#}$)

$$W_s = A_s(I_{1s} - 3I_{3s}^{1/3}) + B_s(I_{2s} - 3I_{3s}^{2/3}) + \frac{\hat{\lambda}_s}{2} \ln^2(I_{3s}^{1/2}) \quad (25)$$

where I_{1s} , I_{2s} , and I_{3s} are the strain invariants for each solid's mappings (equation (1)). The material constants are listed in Table 1.

Table 1. Material constants for solids.

Solid #	A_s	B_s	$\hat{\lambda}_s$	η_s
1	80.	20.	50,000.	-
2	20.	5.	12,500.	2.
3	20.	5.	12,500.	10.

Details on the element mappings, the expressions for the work performed during inflation, etc., that are not provided in Reference¹⁶ are included in Reference¹⁷.

The pressure vs time curve for the first example is given in Figure 7. In this case, the pressure was ramped, held constant, ramped, and then held constant again. Figure 7 also shows the dynamic displacement of the inner surface. The large strain creeping under constant loading is clear. Also, the deformed mesh is compared to the two time dependent reference meshes at both $t = 30$ sec and $t = 75$ sec. Figure 8 shows the pressure vs inner radius displacement. The static solutions of Oden¹⁵ are approached in Figure 8 at the end of the creeping (constant load).

A cyclic sawtooth pressure loading was applied in the second example. The pressure vs time curve is given in Figure 9. The inner radius vs time graph demonstrates softening effects (the displacements continue to grow in each cycle). Also, at $t = 10$ sec the time dependent reference geometries and the deformed mesh are compared. Solid #2's reference shape closely follows the deformed shape (fast relaxing material). That is, it is nearly in phase with the deformed shape. Solid #3 (slow relaxing material) is out of phase with the deformed shape in this example. The pressure vs displacement hysteresis loops are shown in Figure 10 with the long term (quasistatic) solution of Oden.

SUMMARY

An internal solid model for large strain viscoelasticity in which hyperelastic solids with time dependent reference shapes (viscohyperelasticity) are superimposed was reviewed. The time dependent kinematics are driven by the relaxation of hyperelastic forces. Like a Maxwell model in small strain theory, both creep and relaxation behavior can be simulated. Viscohyperelasticity, however, allows for the viscous forces to be dependent on both the large strain history of the deformation and on the current large strain state (through the hyperelastic energy functions selected to drive the kinematics). The inflation of a thick walled cylinder was determined using the finite element method to demonstrate the theory.

REFERENCES

1. Johnson, A. R., Quigley, C. J., Cavallaro, C. and Weight, K. D., A large deformation viscoelastic finite element model for elastomers, in The Mathematics of Finite Elements and Applications VII, Edited by J. R. Whiteman, Academic Press Limited, ISBN 0-12-747257-6, 1991.
2. Johnson, A. R. and Quigley, C. J., A viscohyperelastic Maxwell model for rubber viscoelasticity, Rubber Chemistry and Technology, 65, 137-153, 1992.
3. Johnson, A. R., Quigley, C. J., Young, D. G. and Danik, J. A., Viscohyperelastic modeling of rubber vulcanizates, accepted for publication, Tire Science and Technology, Journal of The Tire Society.
4. Johnson, A. R., Quigley, C. J., Weight, K. D., Cavallaro C. and Cox, D. L., Inflation and deflation of a thick walled viscohyperelastic sphere, The Transactions of the Eighth Army Conference on Applied Mathematics and Computing, U.S. Army Research Office Report No. 91-1, 1991, 847-857.
5. Bernstein, B., Kearsley, E. A. and Zapas, L. J., A study of stress relaxation with finite strain, Rubber Chemistry & Technology, 38, 1965, 76-89.
6. McGuirt, C. W. and Lianis, G., Experimental investigation of non-linear, non-isothermal viscoelasticity, International Journal of Engineering Science, 7, 1969, 579-599.
7. McGuirt, C. W. and Lianis, G., Constitutive equations for viscoelastic solids under finite uniaxial and biaxial deformations, Transactions of the Society of Rheology, 14, 1970, 117-134.
8. Green, A. E. and Rivlin, R. S., The mechanics of nonlinear materials with memory, Archives of Rational Mechanics Analysis, 1, 1965, 76-89.
9. Zienkiewicz, O. C., Watson, M. and King, I. P., A numerical method of visco-elastic stress analysis, International Journal of Mechanical Science, 10, 1968, 807-827.
10. Carpenter, W. C., Viscoelastic stress analysis, International Journal of Numerical Methods in Engineering, 4, 1972, 357-366.
11. Treloar, L. R. G., The Physics of Rubber Elasticity, Clarendon Press, Oxford, 1975.
12. Ogden, R. W., Non-Linear Elastic Deformations, Ellis Horwood Limited, Chichester, 1984.
13. Green, A. E. and Zerna, W., Theoretical Elasticity, Oxford University Press, 1968.
14. Green, M. S. and Tobolsky, A. V., A new approach to the theory of relaxing polymeric media, J. Chem. Phys. 14, 1946, 30-92.

15. Oden, J. T., Finite Elements of Nonlinear Continua, McGraw-Hill Inc., 1972.
16. Fried, I. and Johnson, A. R., A note on the elastic energy density functions for largely deformed compressible rubber solids, *Comput. Meths. Appl. Mech. Engrg.* 69, 1988, 53-64.
17. Johnson, A. R., Quigley, C. J., Freese, C. E. and Cox, D. L., A viscohyperelastic finite element model for rubber, in preparation.

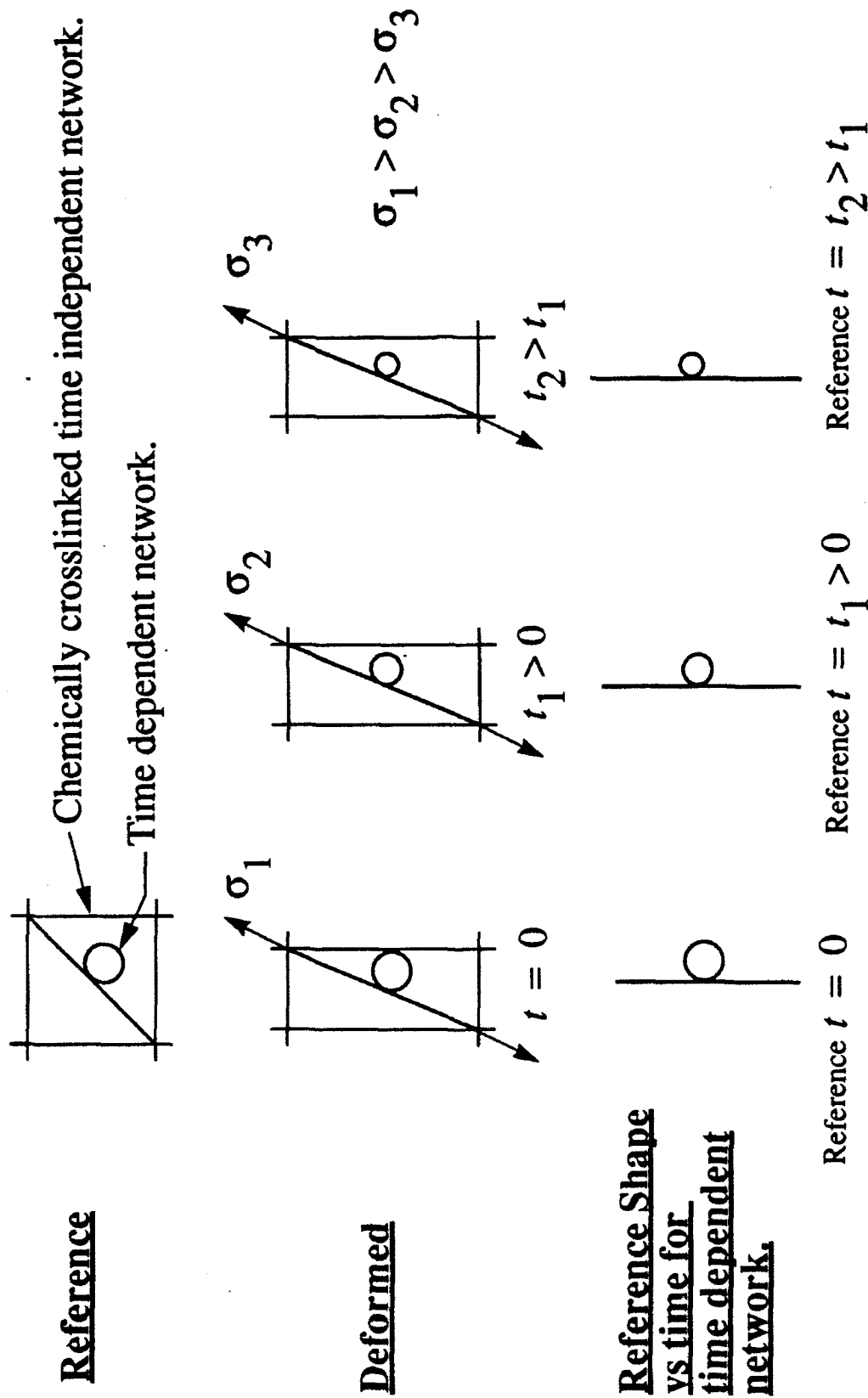


Figure 1. Conceptual model of step strain relaxation.

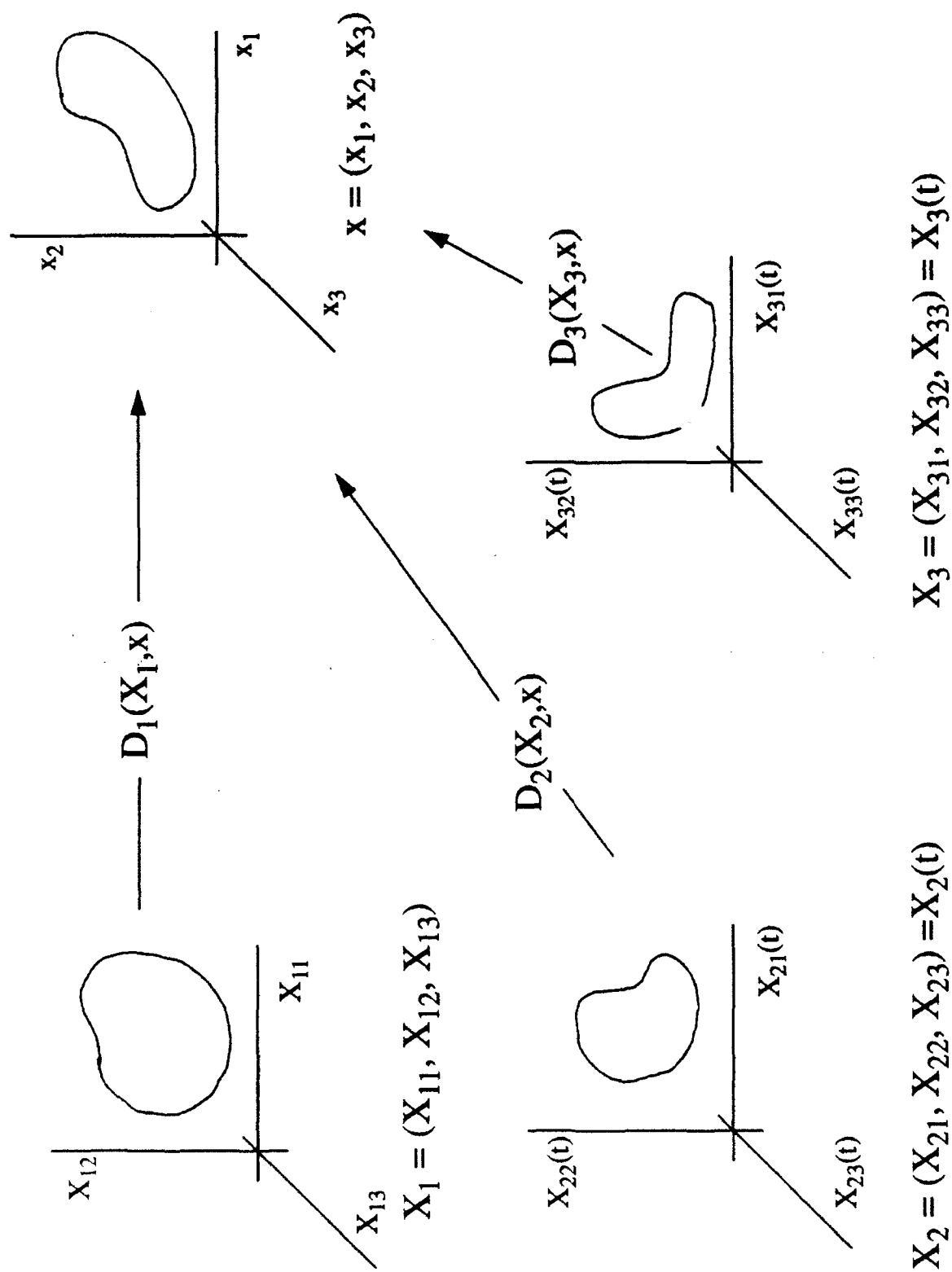
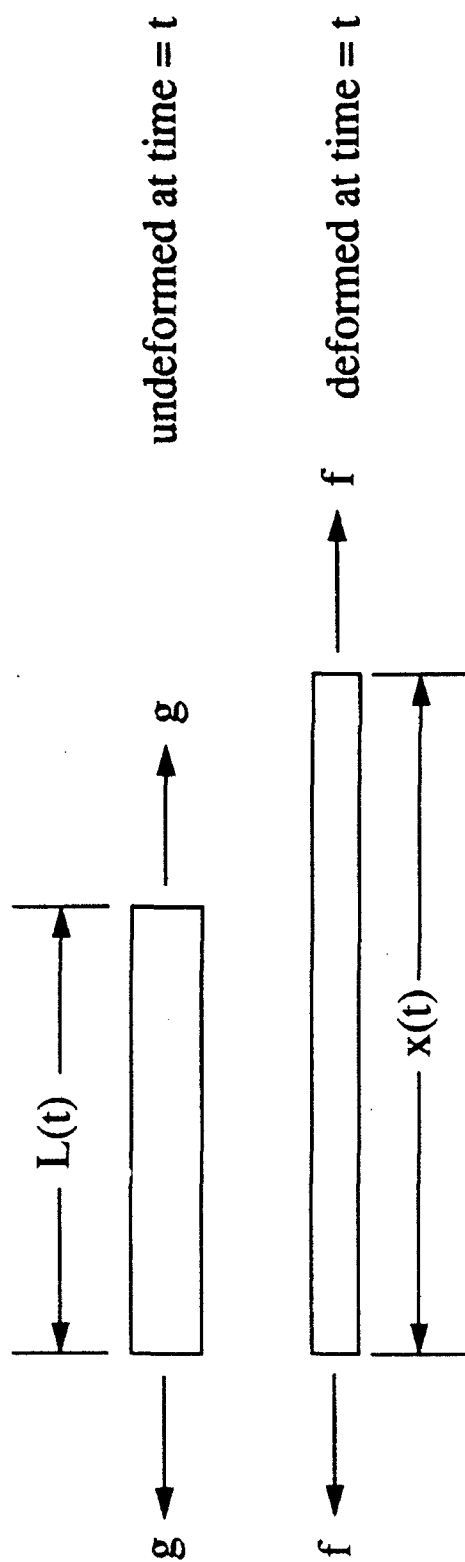


Figure 2. Reference shapes and deformation mappings.



$$\lambda = \frac{x}{L} \qquad f = \mu \left(\lambda - \frac{1}{\lambda^2} \right) \qquad \text{applied force (measured)}$$

$$\Lambda = \frac{L}{x} \qquad g = \mu \left(\Lambda - \frac{1}{\Lambda^2} \right) \qquad \text{flow force}$$

$$(-\eta) \frac{\partial g}{\partial t} = g \qquad \text{flow(relaxation) equation}$$

Figure 3. One dimensional model of time dependent network.

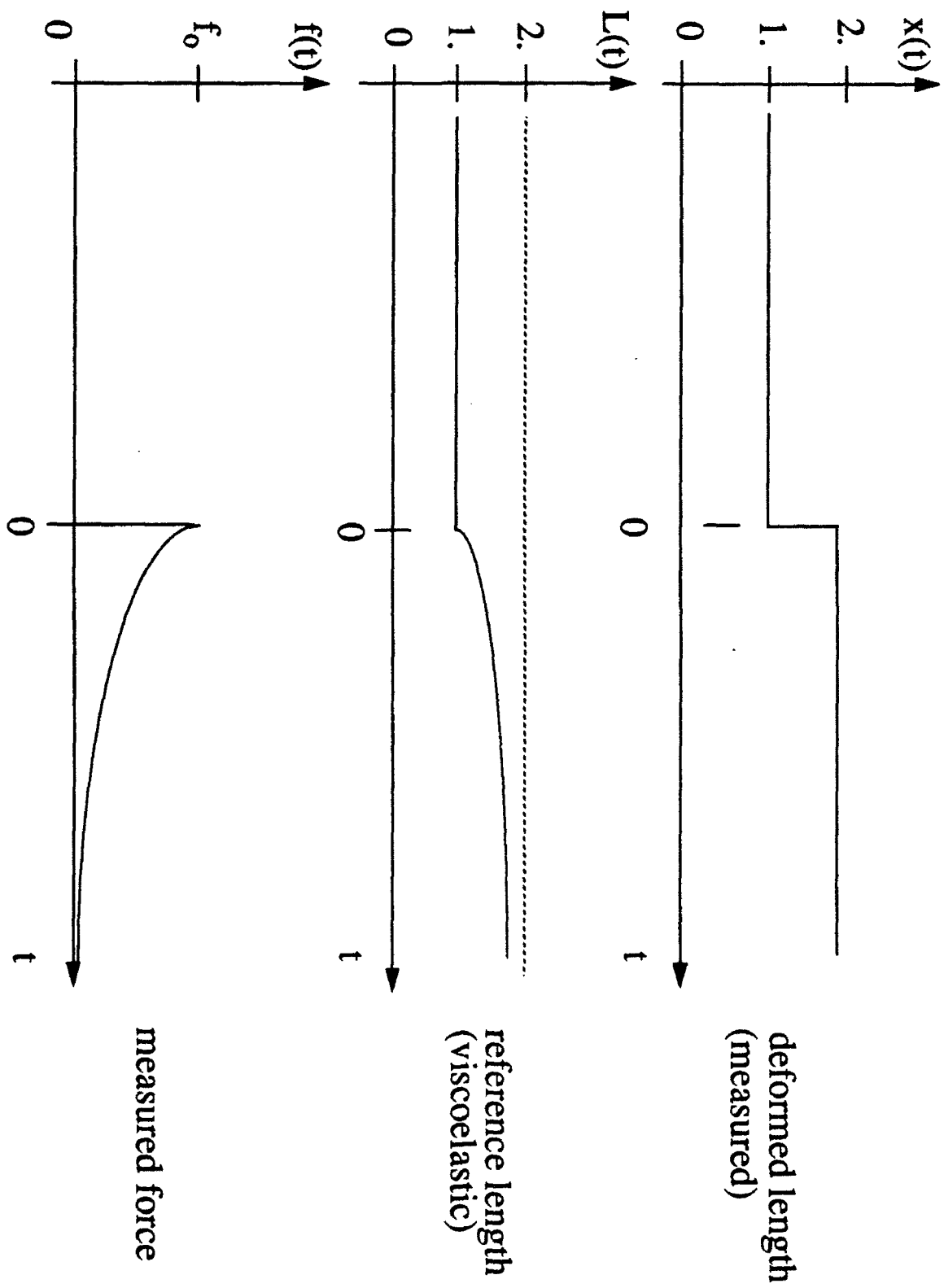


Figure 4. One dimensional step-strain relaxations.

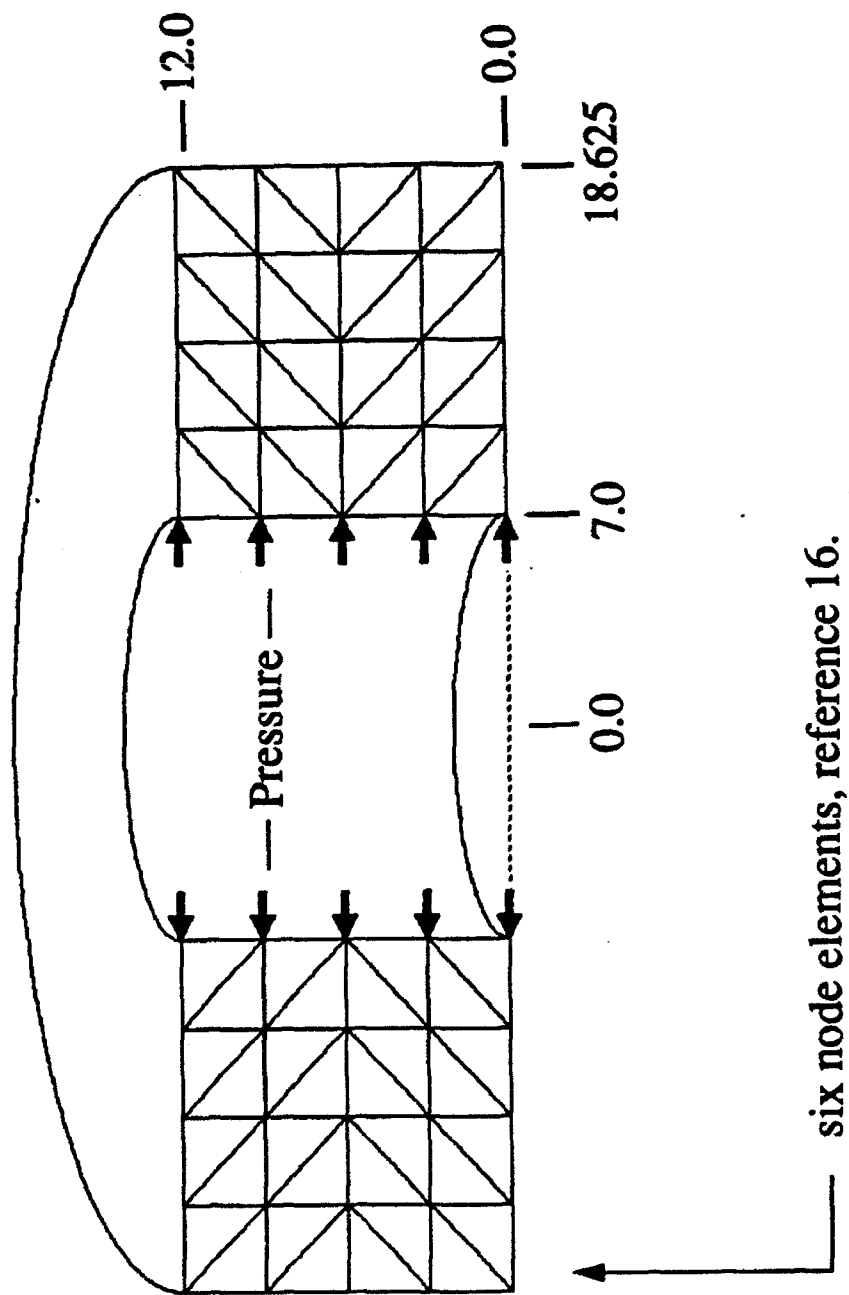


Figure 6. Dynamic internal pressure loading of a thick-walled cylinder.

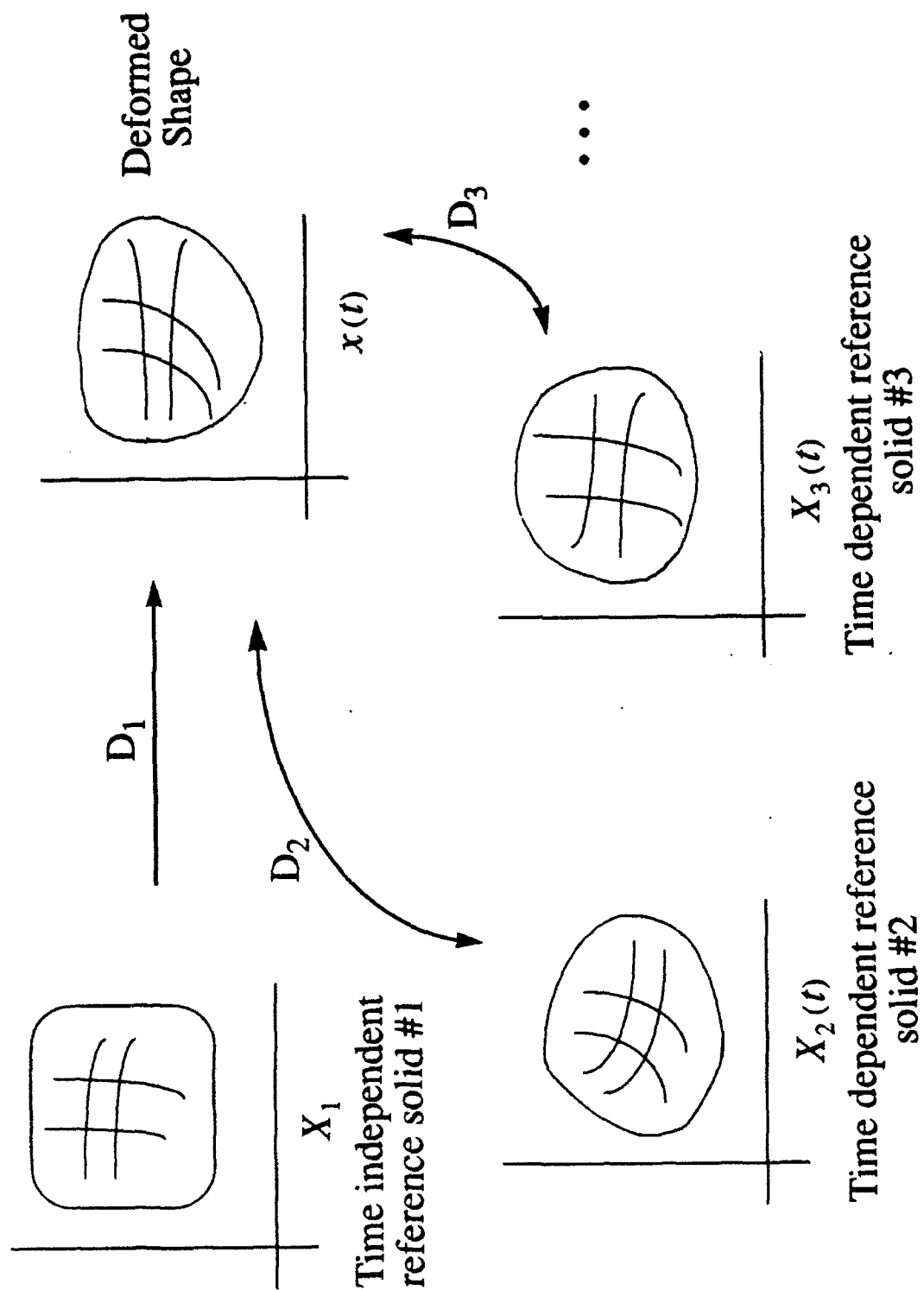
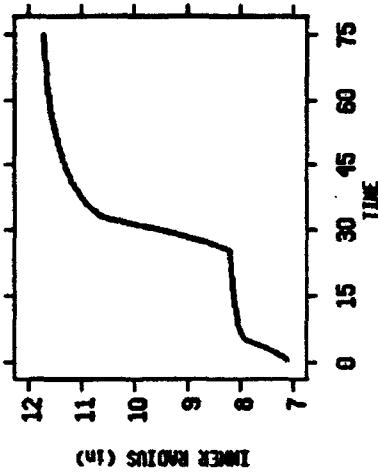
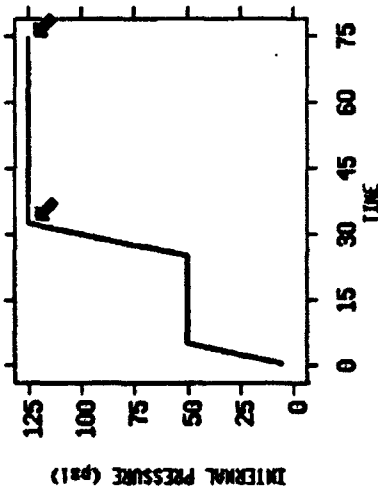
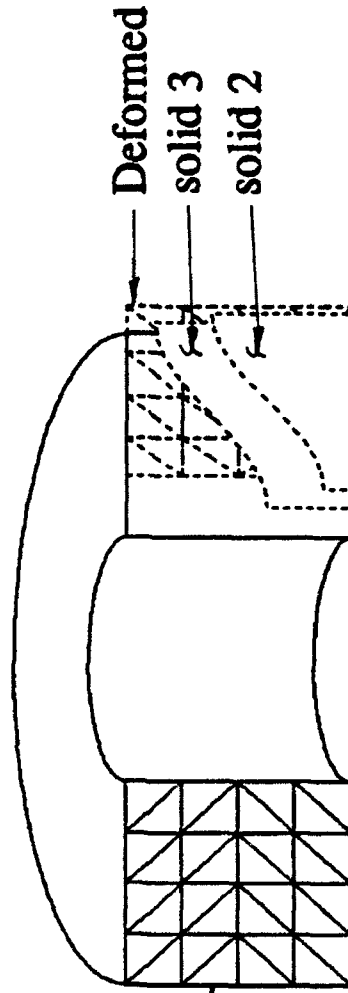


Figure 5. Finite element discretization of reference and deformed shapes.



t = 30 sec



t = 75 sec

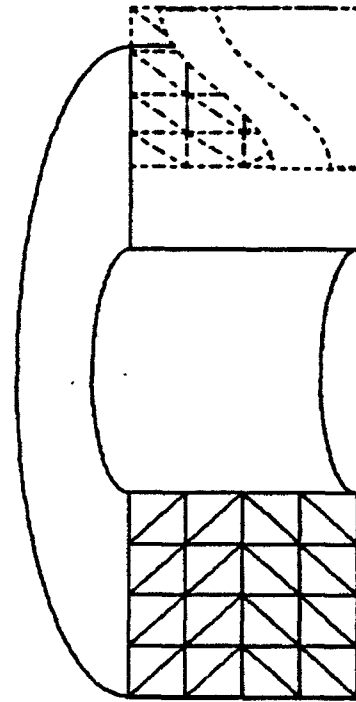


Figure 7. Reference and deformed meshes at t = 30 sec and t = 75 sec.

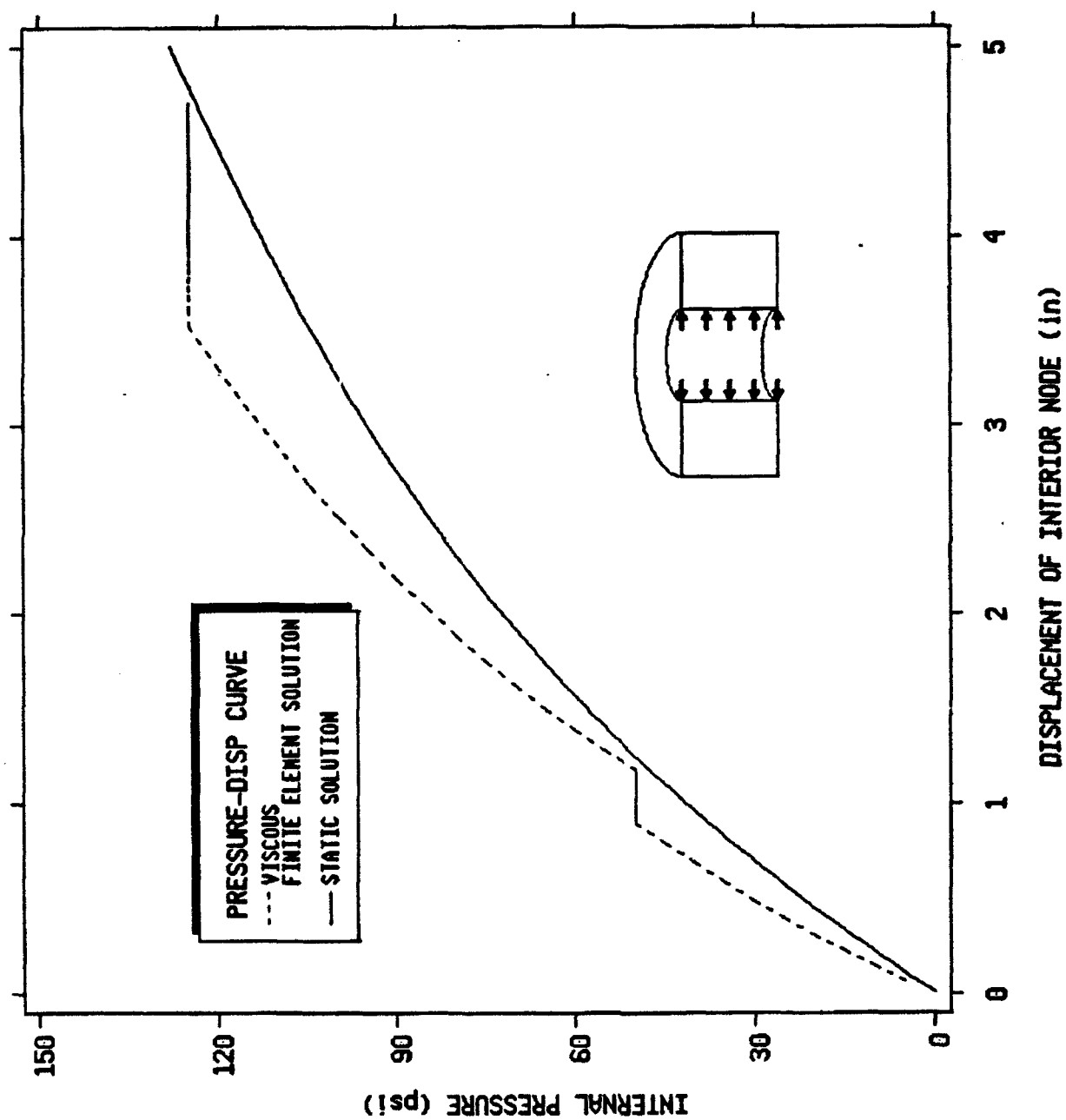


Figure 8. Pressure vs radial displacement of interior surface.

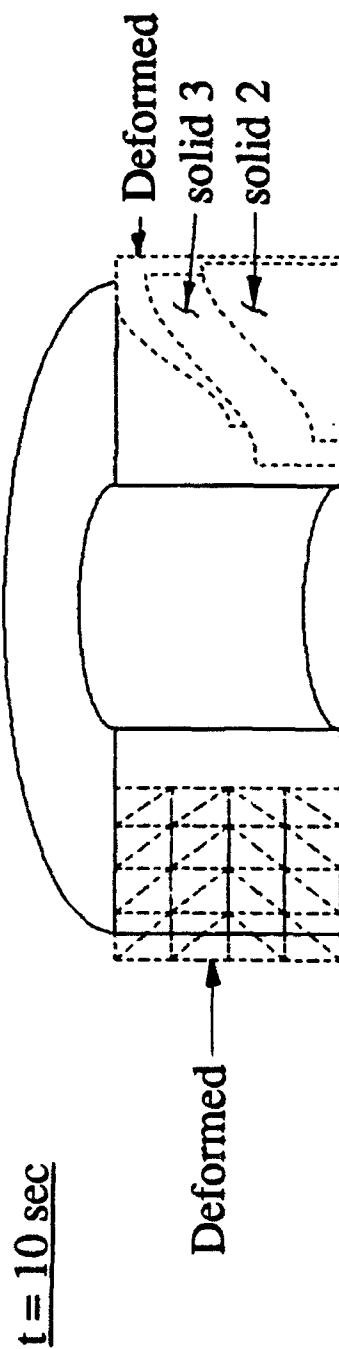
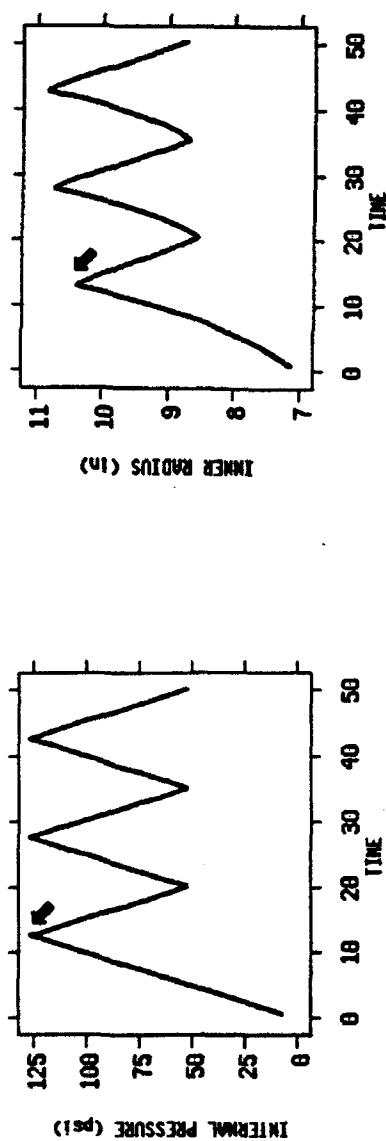


Figure 9. Deformed meshes at $t = 10$ sec, cyclic loading.

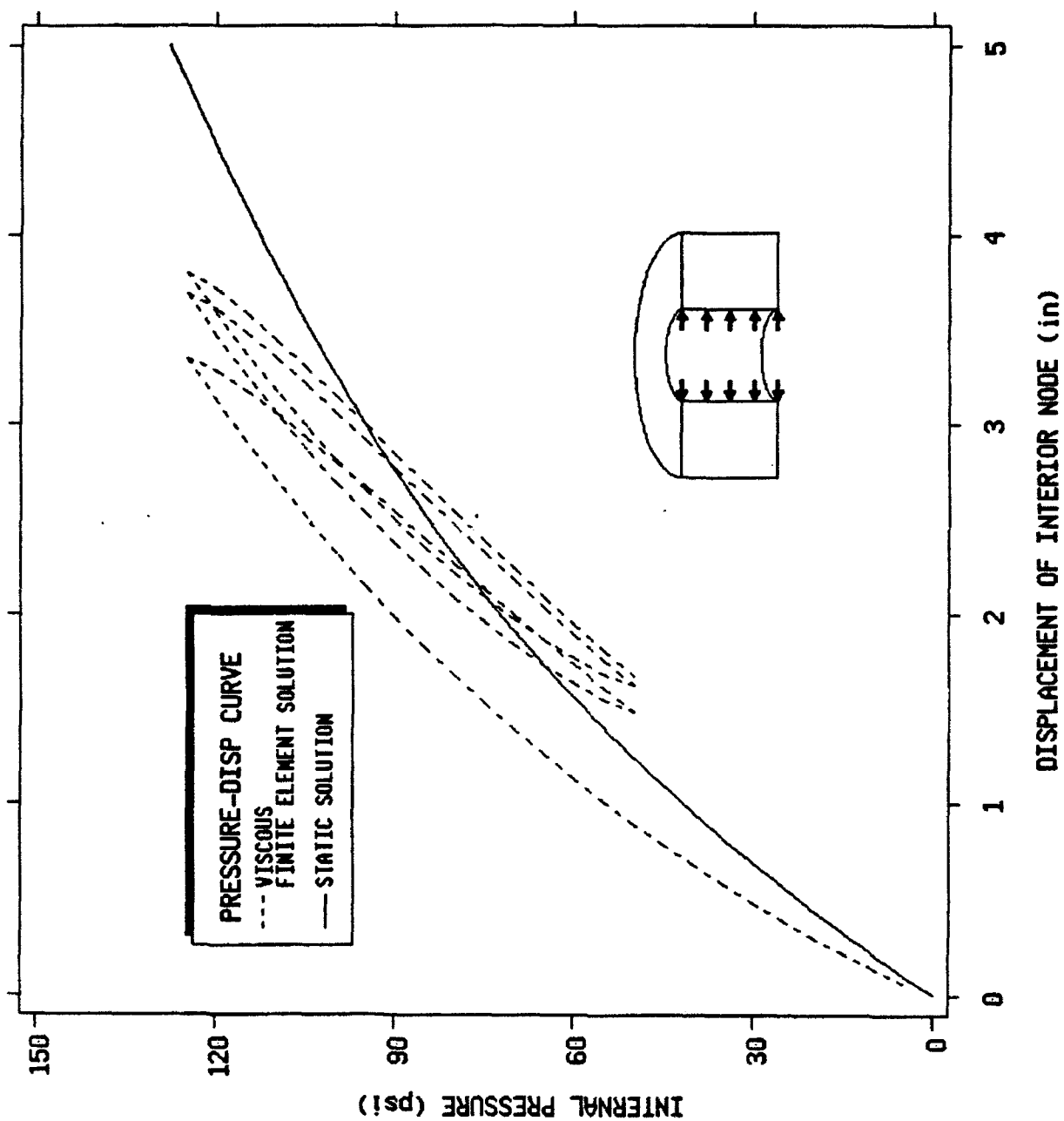


Figure 10. Pressure vs radial displacement of interior surface, cyclic loading.

MULTIGRID ALGORITHMS FOR THE FIRST BIHARMONIC PROBLEM: ROBUSTNESS*

M. R. HANISCH
MATHEMATICAL SCIENCES INSTITUTE
CORNELL UNIVERSITY
ITHACA, NY 14853

Abstract. We consider the numerical solution of the first biharmonic problem. Numerous discretizations proposed for this problem lead to large, sparse, and extremely ill-conditioned linear systems $\mathcal{A}U = F$. We shall describe certain fast solvers for such systems. In particular, for the mixed method of Ciarlet and Raviart, and for the nonconforming finite element method of Morley, we describe multigrid iterative solvers for the systems $\mathcal{A}U = F$.

Theoretical results and practical computations have shown that the multigrid V-cycle and W-cycle iterations work extremely well for second-order elliptic problems. However, the V-cycle and W-cycle iterations may diverge when applied to the first biharmonic problem. It is nevertheless possible to construct provably robust multigrid iterations for these fourth-order problems. From a single "Variable V-cycle" iteration we construct a preconditioner \mathcal{B} with the property that $\mathcal{B}\mathcal{A}$ is uniformly well-conditioned. Furthermore, $\mathcal{B}\mathcal{A}$ is symmetric with respect to an appropriate inner product. The linear systems $\mathcal{A}U = F$ may then be rapidly solved with a preconditioned conjugate gradient iteration. The effectiveness of this strategy will be illustrated computationally. For example, on a 255x255 mesh where the condition number of \mathcal{A} may exceed 10^8 , we obtain $\mathcal{B}\mathcal{A}$ with condition number typically less than 20.

1. Introduction. In this paper we consider the behavior of multigrid iterations applied to finite element schemes for the first biharmonic problem in the plane. The first biharmonic problem models the bending of clamped elastic plates and in fluid dynamics, two-dimensional Stokes flow. We specifically consider the Ciarlet-Raviart mixed method [11] and the nonconforming method of Morley [23]. Additional mixed methods, the method of Herrmann and Miyoshi [16, 17, 22], and of Herrmann and Johnson [16, 17, 18], and the method of Raviart and Thomas [25] for second-order problems, may similarly be considered; cf. [15]. Each of these methods leads one to solve a sparse but extremely ill-conditioned linear system. And in the case of the Ciarlet-Raviart mixed method we obtain an indefinite block matrix equation of the form

$$(1.1) \quad \mathcal{N} \begin{pmatrix} S_h \\ U_h \end{pmatrix} = \begin{pmatrix} \mathcal{Q} & \mathcal{D}^t \\ \mathcal{D} & 0 \end{pmatrix} \begin{pmatrix} S_h \\ U_h \end{pmatrix} = \begin{pmatrix} -G \\ -F \end{pmatrix}.$$

Here U_h and S_h are vectors of approximate nodal values for the solution and its negative Laplacian. For simplicity we shall take $G = 0$, in which case (1.1) corresponds to the Dirichlet problem with homogeneous boundary data. Eliminating S_h from (1.1), we obtain an equation involving an ill-conditioned but typically symmetric positive definite Schur complement

$$(1.2) \quad \mathcal{A}U_h = [\mathcal{D}\mathcal{Q}^{-1}\mathcal{D}^t]U_h = F.$$

With U_h obtained from (1.2), one may then compute $S_h = -\mathcal{Q}^{-1}\mathcal{D}^tU_h$. The costly computation of \mathcal{Q}^{-1} may be avoided using a technique described in Section 2.1.

* This work was partly supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University.

It can be proved that a multigrid W-cycle iteration which uses a sufficiently large but undetermined number of smoothing steps, a number we shall parametrize with m , can be used to solve the Morley equations and also the Ciarlet-Raviart equations through (1.2). The hypothesis that m is sufficiently large is common to many multigrid analyses. For example, Brenner [9] has proved convergence of a W-cycle iteration for the Morley equations when m is sufficiently large. And if m is sufficiently large, a W-cycle iteration for the Ciarlet-Raviart normal equations, $\mathcal{N}^2 = \mathcal{N}^t \mathcal{N}$, can be proved to converge; see [24]. Furthermore, practical computations demonstrate that $m = 1$ is large enough for W-cycle convergence for many problems. However, for the Morley and Ciarlet-Raviart equations, larger values of m may be required for W-cycle convergence. A numerical example is given in Section 4 for which W-cycle convergence is not guaranteed unless $m \geq 8$. As a value for m sufficient for convergence cannot be specified a priori, W-cycle iterations are not robust for these problems.

To obtain a robust multigrid scheme, we will construct a multigrid preconditioner \mathcal{B} for the Schur complement \mathcal{A} . The action of this preconditioner is computed by performing a single iteration of a particular multigrid scheme. A multigrid preconditioner will similarly be constructed for the symmetric positive definite Morley system. Conjugate gradient iterations for the preconditioned problems, for example

$$(1.3) \quad \mathcal{B} \mathcal{A} U_h = \mathcal{B} F,$$

are known to converge with a rate which is bounded by a function of the condition number of $\mathcal{B} \mathcal{A}$, $\kappa_2(\mathcal{B} \mathcal{A})$. The smaller the condition number, the better the bound, as we shall see in (3.20) of Section 3.2. It will be shown that a "Variable V-cycle" multigrid preconditioner yields problems (1.3) with small condition number, bounded independently of the mesh diameter for a scale of underlying finite element meshes. This is an improvement over a result of Braess and Peisker [4].

The lack of W-cycle robustness observed with the Morley and Ciarlet-Raviart methods can be understood to be a consequence of certain features; namely, the nonnestedness of the Morley finite element spaces and the noninheritedness of quadratic forms induced by equations of the form (1.2). However, a multigrid analysis based on the theory of Bramble, Pasciak, and Xu [8] can be provided in each case. Accordingly, an "Approximation and Regularity" property must be proved.

We will assume that problems are posed on convex polygonal domains Ω ; i.e., H^3 -regularity is obtained for the first biharmonic problem. The multigrid analysis is extended to nonconvex polygonal domains in [15]. Finally, we note that in the "inherited form, nested space" setting the W-cycle is provably robust; see [2, 3, 6, 19, 20, 21].

This paper is arranged in the following manner. In Section 2, we outline the various finite element methods considered for the first biharmonic problem. The multigrid preconditioner is described in Section 3. In the final section, we present the results of several computations.

Throughout this paper we use C to denote a generic positive constant which is independent of the mesh parameter h .

2. Finite element methods. Given a convex polygonal domain Ω in \mathbf{R}^2 with boundary $\partial\Omega$, the first biharmonic problem (with homogeneous boundary data) is:

$$(2.1) \quad \begin{aligned} \Delta^2 u &= f & \text{in } \Omega, \\ u &= \frac{\partial u}{\partial \nu} = 0 & \text{on } \partial\Omega, \end{aligned}$$

where Δ denotes the Laplacian operator and $\frac{\partial}{\partial \nu}$ is the normal derivative at the boundary of Ω . This problem provides a simple model for the displacement of a clamped elastic plate, or for the stream function of a steady-state planar Stokes flow.

Recall the definition, for nonnegative integer s , of the Sobolev spaces containing functions with square-integrable derivatives,

$$(2.2) \quad H^s(\Omega) \stackrel{\text{def}}{=} \{v \in L_2(\Omega) : D^\alpha v \in L_2(\Omega), \text{ for } |\alpha| \leq s\},$$

where D denotes the distributional or weak derivative, and α is a multi-index. The Sobolev space $H^s(\Omega)$ is a Hilbert space and we shall denote its norm by $\|v\|_s$. We will refer to additional spaces $H_0^s(\Omega)$ which may be defined as the completions of $C_0^\infty(\Omega)$ with respect to the norms $\|\cdot\|_s$. (Denote by $C_0^\infty(\Omega)$ the space of infinitely differentiable functions with compact support contained in Ω .) A useful a priori inequality can be obtained for solutions to (2.1) when f is given in certain negative norm (dual) spaces $H^{-s}(\Omega) \stackrel{\text{def}}{=} [H_0^s(\Omega)]'$. In particular, with Ω a convex polygonal domain and $f \in H^{-1}(\Omega)$, the unique existence of a solution $u \in H^3(\Omega) \cap H_0^2(\Omega)$ to (2.1) is known and we have

$$(2.3) \quad \|u\|_3 \leq C \|f\|_{-1},$$

for a constant C which is independent of f , see [13].

2.1. The Ciarlet-Raviart method. Introducing an auxiliary variable σ such that

$$\sigma = -\Delta u \quad \text{and} \quad -\Delta \sigma = f,$$

and using a standard Green's formula, one may obtain a weak formulation for (2.1).

$$(2.4) \quad \left\{ \begin{array}{l} \text{Find: } \{\sigma, u\} \in H^1(\Omega) \times H_0^1(\Omega) \text{ such that for } f \in H^{-1}(\Omega), \\ (\sigma, v)_{L_2} - D(v, u) = 0 \quad \forall v \in H^1(\Omega), \\ -D(\sigma, w) = -(f, w) \quad \forall w \in H_0^1(\Omega), \end{array} \right.$$

where the "Dirichlet form", $D(\cdot, \cdot)$, and the L_2 -inner product are given by

$$D(\varphi, v) \stackrel{\text{def}}{=} \int_{\Omega} \nabla \varphi \cdot \nabla v \, dx \leq \|\varphi\|_1 \|v\|_1, \quad (v, \varphi)_{L_2} \stackrel{\text{def}}{=} \int_{\Omega} v \varphi \, dx.$$

This formulation was studied by Ciarlet and Raviart in [11]. It is not difficult to show that the problem (2.4) has a unique solution $\{\sigma, u\} \in H^1(\Omega) \times H_0^1(\Omega)$ for all $f \in H^{-1}(\Omega)$. Furthermore, $\sigma = -\Delta u$, and this same u solves (2.1) in an appropriate sense.

Given a regular and quasi-uniform triangulation (in the sense of Ciarlet [10]) τ_h of Ω , with mesh diameter h , define finite dimensional subspaces $V_h \subset H^1(\Omega)$ and $M_h \subset H_0^1(\Omega)$ from the space

$$(2.5) \quad S_h^m \stackrel{\text{def}}{=} \{v \in C^0(\bar{\Omega}) : v|_T \in P^{(m)}(T), \quad \forall T \in \tau_h\}.$$

where $P^{(m)}(T)$ is the space of polynomials of degree m or less over triangle T . Set $V_h = S_h^m$ and $M_h = S_h^m \cap H_0^1(\Omega)$. The Ciarlet-Raviart mixed method approximates the solution $\{\sigma, u\}$ of (2.4) with the unique solution to the following problem.

$$(2.6) \quad \left\{ \begin{array}{l} \text{FIND: } \{\sigma_h, u_h\} \in V_h \times M_h \text{ such that for } f \in H^{-1}(\Omega), \\ (\sigma_h, v)_{L_2} - D(v, u_h) = 0 \quad \forall v \in V_h, \\ -D(\sigma_h, w) = -(f, w) \quad \forall w \in M_h. \end{array} \right.$$

Choosing bases $\{\phi^i\}$ for V_h and $\{\phi^j\}$ for M_h , consider the linear system (2.6) in block matrix form with the notations $[\mathcal{D}_h]_{ij} = -D(\phi^j, \phi^i)$, $[\mathcal{Q}_h]_{ij} = (\phi^i, \phi^j)_{L_2}$, $[F]_j = (f, \phi^j)$, and denote the transpose of \mathcal{D}_h by \mathcal{D}_h^t . Applying block Gaussian elimination we obtain the reduced system for (the coefficients of) u_h

$$(2.7) \quad \mathcal{D}_h \mathcal{Q}_h^{-1} \mathcal{D}_h^t U_h = F.$$

As was noted in the Introduction, it is possible to avoid computing the action of the inverted Gramm matrix \mathcal{Q}_h^{-1} . In particular, we replace the L_2 inner product with a certain approximating bilinear form $(\cdot, \cdot)_h$. Using the approximate form leads to the new problem.

$$(2.8) \quad \left\{ \begin{array}{l} \text{FIND: } \{\sigma_{d,h}, u_{d,h}\} \in V_h \times M_h \text{ such that for } f \in H^{-1}(\Omega), \\ (\sigma_{d,h}, v)_h - D(v, u_{d,h}) = 0 \quad \forall v \in V_h, \\ -D(\sigma_{d,h}, w) = -(f, w) \quad \forall w \in M_h. \end{array} \right.$$

We then consider the associated reduced system, perhaps using different computational bases

$$(2.9) \quad \mathcal{D}_{d,h} \mathcal{Q}_{d,h}^{-1} \mathcal{D}_{d,h}^t U_{d,h} = F,$$

where $\mathcal{Q}_{d,h}$ is the Gramm matrix associated with $(\cdot, \cdot)_h$. For quadratic Ciarlet-Raviart spaces we take

$$(2.10) \quad (u, v)_h = \sum_{T \in \tau_h} \sum_{\alpha=1}^6 \omega_{\alpha,T} \mathcal{F}_{\alpha,T}(u) \mathcal{F}_{\alpha,T}(v),$$

with α indexing the nodes and the midpoints of the edges of triangle T . When α selects a node n_α we use a weight $\omega_{\alpha,T} = \frac{1}{6} \text{area}(T)$ and define the functional $\mathcal{F}_{\alpha,T}(u) = u(n_\alpha)$. If α indexes a midpoint on an edge of T which joins node $n_{\alpha+}$ to node $n_{\alpha-}$, then we set $\omega_{\alpha,T} = \frac{2}{3} \text{area}(T)$ and $\mathcal{F}_{\alpha,T}(u) = u(n_\alpha) - \frac{1}{4}(u(n_{\alpha+}) + u(n_{\alpha-}))$. It is shown in [14] that the problem (2.8) has a unique solution with approximation properties comparable to those of $\{\sigma_h, u_h\}$, the Ciarlet-Raviart approximate solution. Specifically, we have

$$\|u - u_{d,h}\|_1 \leq Ch^2 \|u\|_3 \quad \text{and} \quad \|\sigma - \sigma_{d,h}\|_0 \leq Ch \|u\|_3.$$

Furthermore, in [14] a basis $\{\tilde{\phi}^i\}$ for the space V_h is constructed for which $[\mathcal{D}_{d,h}]_{ij} = -D(\tilde{\phi}^j, \phi^i)$ is sparse and $[\mathcal{Q}_{d,h}]_{ij} = (\tilde{\phi}^i, \tilde{\phi}^j)_h$ is diagonal, and hence is trivially inverted.

2.2. The Morley method. A natural weak formulation of (2.1) seeks a function $u \in H_0^2(\Omega)$ such that

$$(2.11) \quad \int_{\Omega} \Delta u \Delta v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in H_0^2(\Omega).$$

To obtain a so-called conforming finite element method from this weak formulation requires finite element spaces which are contained in $H_0^2(\Omega)$ —this leads to the use of continuously differentiable, hence complex, piecewise polynomial spaces.

An alternative nonconforming finite element method has been proposed by Morley [23]. Let τ_h denote a regular and quasi-uniform triangulation of the domain Ω . The spaces of Morley are defined so that $v \in M_h$ if and only if:

- (a) v restricted to each triangle $T \in \tau_h$ is a quadratic polynomial,
- (b) v is continuous at triangle vertices and vanishes at boundary vertices, and
- (c) the normal derivative $\frac{\partial v}{\partial n}$ is continuous at triangle edge midpoints and vanishes at midpoints along $\partial\Omega$.

The Morley method approximates the solution of (2.11) with the solution to the following problem.

$$(2.12) \quad \begin{cases} \text{FIND: } u_h \in M_h \text{ such that for } f \in L_2(\Omega) \\ \sum_{T \in \tau_h} \int_T \Delta u_h \Delta v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in M_h. \end{cases}$$

Note that the elements of M_h are not continuous at τ_h edge midpoints. It is not difficult to see that if $\tau_{h/2}$ is constructed by joining edge midpoints in τ_h , then M_h is not contained in $M_{h/2}$. Consequently, multigrid methods for Morley discretizations involve nonnested spaces.

It can be shown that $\sum_{T \in \tau_h} \int_T \Delta(\cdot) \Delta(\cdot) \, dx$ is an inner product for the space M_h . Consequently, (2.12) has a unique solution. The Morley approximation u_h is known to satisfy

$$\|u - u_h\|_{2,h} \leq C h [\|u\|_3 + h \|f\|_0],$$

where u denotes the solution of (2.1) (or (2.11)), and $\|v\|_{2,h} \equiv (\sum_{T \in \tau_h} \int_T \Delta v \Delta v \, dx)^{1/2}$.

Since M_h contains discontinuous elements, the requirement $f \in L_2(\Omega)$ is necessary for $\int_{\Omega} f v \, dx$ in (2.12) to make sense. Let v^I denote the piecewise τ_h -linear interpolant of $v \in M_h$. Then the Morley method may be extended to allow $f \in H^{-1}$, (since $v^I \in H_0^1(\Omega)$) as follows,

$$\begin{cases} \text{FIND: } \bar{u}_h \in M_h \text{ such that for } f \in H^{-1} \\ \sum_{T \in \tau_h} \int_T \Delta \bar{u}_h \Delta v \, dx = \int_{\Omega} f v^I \, dx \quad \forall v \in M_h. \end{cases}$$

The unique solution \bar{u}_h of this problem is known to satisfy the error estimate [1]

$$\|u - \bar{u}_h\|_{2,h} \leq C h \|u\|_3.$$

Finally, let $\{\phi^j\}$ denote a computational basis for the space M_h . We may write (2.12) in matrix form

$$\mathcal{M}_h U_h = F_h,$$

where U_h is the coefficient vector for u_h and $[F_h]_i = \int_{\Omega} f \phi^i \, dx$.

3. Multigrid Algorithms. In this section we will describe several multigrid approaches and apply them to the finite element methods of the previous section. In particular, we shall describe a multigrid W-cycle iteration which has as a parameter m the number of (symmetrically applied) smoothing iterations computed during each iteration. As this standard W-cycle will diverge unless m is larger than an (a priori) undetermined value, we propose a second, more robust, multigrid approach. We show that an effective multigrid *preconditioner* can be constructed, and that the action of this preconditioner can be computed as the outcome of an ordinary "Variable V-cycle" multigrid iterative step. We shall also show that the W-cycle is itself a most simple, in this case a too simple multigrid preconditioned iteration.

3.1. Overview. Consider a nested sequence of quasi-uniform (in the sense of Ciarlet [10]) triangulations of a domain Ω with mesh diameters $\{h_k\}_{k=1,\dots,j}$ satisfying a growth condition

$$(3.1) \quad h_{k-1} \leq \theta h_k,$$

and with θ independent of k . Beginning with a coarse triangulation τ_{h_1} of Ω , such a sequence may be obtained by joining the midpoints of the edges of mesh $\tau_{h_{k-1}}$ to form mesh τ_{h_k} . On each mesh we construct a Ciarlet-Raviart or a Morley space M_{h_k} and consider the associated finite element methods. (In the sequel, subscripts h_k will be replaced by the subscript k .) Given local bases, denoted by $\{\phi_k^i\}$, these methods lead to linear systems

$$(3.2) \quad \mathcal{A}_k U_k = F_k.$$

Here, \mathcal{A}_k denotes either the matrix $\mathcal{D}_{d,k} \mathcal{Q}_{d,k}^{-1} \mathcal{D}_{d,k}^t$ or \mathcal{M}_k of Section 2, and U_k denotes the coefficient vector of the approximate solution $u_{d,k}$, or u_k , for the Ciarlet-Raviart, or for the Morley method, respectively. We wish to obtain for $k = j$ the solution to problem (3.2)—the solution on the finest mesh.

Multigrid analyses commonly study quadratic forms associated with the matrices \mathcal{A}_k . Let U and V denote coefficient vectors of u and v , elements of M_k (i.e., $u = \sum_i [U]_i \phi_k^i$). We define

$$A_k(u, v) \equiv \langle \mathcal{A}_k U, V \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. Let $(\cdot, \cdot)_k$ denote inner products for the spaces M_k for which the induced norms $\|u\|_{0,k} = (u, u)_k^{1/2}$ are uniformly (with respect to k) equivalent to the $L_2(\Omega)$ norm. The inner products $(\cdot, \cdot)_k$ could be L_2 inner products and need not be related to the approximate L_2 inner products mentioned in Section 2.1. For each space M_k , let $A_k : M_k \rightarrow M_k$ denote the symmetric (self-adjoint with respect to the inner product $(\cdot, \cdot)_k$) positive definite operator satisfying

$$A_k(u, v) = (A_k u, v)_k \quad \forall u, v \in M_k.$$

(The self-adjointness of A_k is a consequence of the symmetry of \mathcal{A}_k .) With $f_k \in M_k$ defined so that $(f_k, v)_k = \langle F_k, V \rangle = \int_{\Omega} f_k v dx$ for all $v \in M_k$, we may rewrite the equations (3.2) in an equivalent operator form

$$(3.3) \quad A_k u_k = f_k.$$

REMARK 1. The quadratic forms obtained for the Ciarlet-Raviart method have the following useful representation,

$$\begin{aligned}
 A_k(u, u)^{1/2} &= \langle \mathcal{D}_{d,k} \mathcal{Q}_{d,k}^{-1} \mathcal{D}_{d,k}^t U, U \rangle^{1/2} = \langle \mathcal{Q}_{d,k}^{-1} \mathcal{D}_{d,k}^t U, \mathcal{Q}_{d,k} \mathcal{Q}_{d,k}^{-1} \mathcal{D}_{d,k}^t U \rangle^{1/2} \\
 (3.4) \quad &= \sup_{v \in V_k \setminus \{0\}} \frac{\langle \mathcal{Q}_{d,k}(\mathcal{Q}_{d,k}^{-1} \mathcal{D}_{d,k}^t U), V \rangle}{\langle \mathcal{Q}_{d,k} V, V \rangle^{1/2}} \\
 &= \sup_{v \in V_k \setminus \{0\}} \frac{D(v, u)}{(v, v)_{h_k}}^{1/2},
 \end{aligned}$$

making use of the fact that $\langle \mathcal{Q}_{d,k} \cdot, \cdot \rangle$ induces an inner product for the space V_k . It is not surprising then that these forms are "noninherited", i.e., that for $k < j$ there is some $u \in M_k \subset M_j$ such that $A_k(u, u) \neq A_j(u, u)$. \square

To define a multigrid algorithm, assume further that one has linear mappings, "prolongations", $I_k : M_{k-1} \rightarrow M_k$. For example, if the space M_{k-1} is contained in M_k , then I_k may be taken to be the natural injection operator. See [9] for a suitable I_k in the Morley setting. Let \mathcal{I}_k denote the matrix for I_k given in terms of the computational bases for M_k and M_{k-1} . It can be shown that the operator $P_{k-1}^o : M_k \rightarrow M_{k-1}$ induced by the transposed matrix \mathcal{I}_k^t satisfies

$$(3.5) \quad (P_{k-1}^o u, v)_{k-1} = (u, I_k v)_k \quad \forall v \in M_{k-1}.$$

Finally, we shall employ linear "smoothing iterations" associated with the problems (3.2), for $k = 2, \dots, j$. We postpone the discussion of the suitability of a smoothing iteration, but now claim that point, line, or block Jacobi or Gauss-Seidel iterations, or the Richardson iteration, may be effectively used. These smoothing iterations can be expressed in terms of procedures

$$X^i = \mathcal{R}_k(X^{i-1}, F_k),$$

where each $\mathcal{R}_k(\cdot, \cdot)$ satisfies

- (1) (consistency) $X = \mathcal{R}_k(X, \mathcal{A}_k X)$ for all X , and
- (2) (linearity) $\mathcal{R}_k(X, Y) + \alpha \mathcal{R}_k(U, V) = \mathcal{R}_k(X + \alpha U, Y + \alpha V)$.

Alternatively, if we define a *matrix* \mathcal{R}_k so that $\mathcal{R}_k X = \mathcal{R}_k(0, X)$ for all X , then we may rewrite the smoothing iteration. Combining properties (1) and (2) we have

$$X^{i-1} = \mathcal{R}_k(0, \mathcal{A}_k X^{i-1}) + \mathcal{R}_k(X^{i-1}, 0)$$

hence

$$\begin{aligned}
 X^i &= \mathcal{R}_k(X^{i-1}, F_k) \\
 &= \mathcal{R}_k(X^{i-1}, 0) + \mathcal{R}_k(0, F_k) \\
 &= X^{i-1} + \mathcal{R}_k(0, F_k - \mathcal{A}_k X^{i-1}) \\
 (3.6) \quad X^i &= X^{i-1} + \mathcal{R}_k[F_k - \mathcal{A}_k X^{i-1}].
 \end{aligned}$$

The smoothing procedures $Z = \mathcal{R}_k(X, F)$ induce analogous procedures $z = R_k(x, f)$ acting directly on elements of the space M_k . Here X and Z denote coefficient vectors of $x \in M_k$

and $z \in M_k$ and f is related to the i th component of F by $[F]_i = (f, \phi_k^i)_k$. Consistency, $x = R_k(x, A_k x)$, and the linearity of $R_k(\cdot, \cdot)$ follow from the same properties of $\mathcal{R}_k(\cdot, \cdot)$. From $R_k(\cdot, \cdot)$ we obtain a smoothing operator $R_k: M_k \rightarrow M_k$, $R_k x \equiv R_k(0, x)$, and we can write the smoothing iteration (3.6) in operator form

$$(3.7) \quad x^i = x^{i-1} + R_k(f_k - A_k x^{i-1})$$

where f_k satisfies $(f_k, v)_k = (F_k, V)$.

REMARK 2. In order that we may later symmetrize our multigrid schemes, we allow for a somewhat more general smoothing; cf. [8]. For example, if R_k denotes the (asymmetric) smoothing operator induced by one particular Gauss-Seidel sweep, then R_k^t is the operator which corresponds to a Gauss-Seidel sweep in the "reverse" direction. Here, the superscript t denotes the adjoint or transpose with respect to $(\cdot, \cdot)_k$. We shall define a Gauss-Seidel operator $R_k^{(i)}$ so that

$$R_k^{(i)} = \begin{cases} R_k & \text{if } i \text{ is odd,} \\ R_k^t & \text{if } i \text{ is even,} \end{cases}$$

and consider smoothing iterations

$$(3.8) \quad x^i = x^{i-1} + R_k^{(i)}(f_k - A_k x^{i-1}).$$

We shall use the notation $X^i = \mathcal{R}_k^{(i)}(X^{i-1}, F_k)$ to denote the action of this alternating smoother. \square

We may now define a rather general symmetric multigrid process for iteratively solving (3.2). Given an initial approximation Z_k^{l-1} to the solution U_k of the problem $A_k U_k = F_k$, compute an improved approximation, $Z_k^l = \text{Mg}_k(Z_k^{l-1}, F_k)$. The procedure $\text{Mg}_k(\cdot, \cdot)$ is defined below by the recursive Algorithm 1. Setting $p = 2$ in this algorithm and using $m(k) = m$ smoothings for each level k yields a multigrid W-cycle. With $p = 1$, a V-cycle is obtained. It is possible to increase the number of smoothings $m(k)$ as k decreases without significantly increasing the cost required to compute $\text{Mg}_k(\cdot, \cdot)$. We shall refer to a Variable V-cycle as that scheme obtained from Algorithm 1 with $p = 1$ and

$$\beta_0 m(k) \leq m(k-1) \leq \beta_1 m(k) \quad \text{and} \quad 1 < \beta_0 \leq \beta_1.$$

Observe that the Variable V-cycle with $m(k) = 2^{j-k}$ and the W-cycle with $m = 1$ require the same number of smoothing iterations for each level, and hence require roughly the same computational effort to perform.

It can be shown that the multigrid procedure is linear and is also consistent with $X = \text{Mg}_k(X, A_k X)$. Consider again the multigrid solution of (3.2) which generates iterates

$$(3.9) \quad Z^l = \text{Mg}_j(Z^{l-1}, F_j)$$

from an initial guess Z^0 . Repeating the arguments used to obtain (3.6) and defining matrices B_k so that $B_k X = \text{Mg}_k(0, X)$, we may rewrite (3.9) in the form

$$(3.10) \quad Z^l = Z^{l-1} + B_j(F_j - A_j Z^{l-1}).$$

procedure $\text{Mg}_k(Z_k^{l-1}, F_k)$

if $k = 1$, solve exactly

return $\text{Mg}_1(Z_1^{l-1}, F_1) = A_1^{-1} F_1$

else, define $\text{Mg}_k(Z_k^{l-1}, F_k)$ in terms of $\text{Mg}_{k-1}(\cdot, \cdot)$ as follows:

0. initialize $X^0 = Z_k^{l-1}$ and Q^0

1. smooth $m(k)$ times

$$X^i = X^{i-1} + \mathcal{R}_k^{(i)}(F_k - A_k X^{i-1}), \quad i = 1, \dots, m(k)$$

2. perform a coarse grid correction

$$Y^{m(k)} = X^{m(k)} + \mathcal{I}_k Q^p$$

where

$$Q^j = \text{Mg}_{k-1}(Q^{j-1}, \mathcal{I}_k^t(F_k - A_k X^{m(k)})), \quad j = 1, \dots, p$$

3. smooth $m(k)$ more times

$$Y^i = Y^{i-1} + \mathcal{R}_k^{(i)}(F_k - A_k Y^{i-1}), \quad i = m(k) + 1, \dots, 2m(k)$$

return $\text{Mg}_k(Z_k^{l-1}, F_k) = Y^{2m(k)}$

ALGORITHM 1. A symmetric multigrid procedure.

or in an equivalent operator form

$$(3.11) \quad z^l = z^{l-1} + B_j(f_j - A_j z^{l-1}).$$

We conclude that the multigrid iteration is simply a linear iterative scheme for solving the preconditioned system

$$(3.12) \quad B_j A_j U_j = B_j F_j \quad \text{or} \quad B_j A_j u_j = B_j f_j.$$

Denoting the l^{th} error $e^l = u_j - z^l$, from (3.11) it is clear that

$$e^l = E_j e^{l-1}$$

holds for a linear error reduction operator $E_j : M_j \rightarrow M_j$. In fact,

$$(3.13) \quad E_j = I - B_j A_j.$$

If the smoothing iterations are "symmetrically performed", see Remark 2, then B_j can be shown to be symmetric with respect to $(\cdot, \cdot)_j$. Consequently, E_j is symmetric with respect to the inner product $(A_j \cdot, \cdot)_j$. The multigrid iteration (3.9) is contracting provided that the eigenvalues of the operator E_j are contained in the interval $(-1, 1)$, or if $\|E_j\| \leq \delta < 1$.

Equivalently, the linear iterative scheme (3.11) converges provided that the eigenvalues of $B_j A_j$ are contained in $(0, 2)$. For a result in this direction, see Theorem 4.

If the maximum eigenvalue of $B_j A_j$ is larger than 2, then (3.11) generally diverges; although, a different iterative scheme for solving (3.12), conjugate gradients for example, may rapidly converge. This is an important observation for the multigrid iteration when it is applied to the finite element methods of Section 2. Indeed, the W-cycle with $m = 1$ may diverge for these methods. However, the problem (3.12) with B_j obtained from the Variable V-cycle with $p = 1$ and $m(k) = 2^{j-k}$ has a small condition number—independent of the mesh diameter h_k —and is rapidly solved by a conjugate gradient iteration; cf. Theorem 5. Even when $\|E_j\| \leq \delta < 1$ so that the multigrid iteration (3.11) converges, the spectral condition number of $B_j A_j$ satisfies

$$\kappa_2(B_j A_j) = \kappa_2(B_j A_j) = \frac{\lambda_{\max}(B_j A_j)}{\lambda_{\min}(B_j A_j)} \leq \frac{1 + \delta}{1 - \delta}$$

and it is often faster to solve (3.12) with a conjugate gradient iteration.

Using the inner product $[\cdot, \cdot] = \langle B_j^{-1} \cdot, \cdot \rangle$ in which $B_j A_j$ is symmetric, a conjugate gradient algorithm for $B_j A_j U_j = B_j F_j$ can be derived; see Algorithm 2. (Note that the matrix $B_j A_j$ may not be symmetric.) Algorithm 2 happens to be identical to the preconditioned conjugate gradient algorithm for the problem $B_j^{1/2} A_j B_j^{1/2} (B_j^{-1/2} U_j) = B_j^{1/2} F_j$ described in [12].

```

l = 0;   X0 = 0;   R0 = Fj
while   ||Zl|| > ε ||Z0||
  compute Zl = Bj Rl
  l = l + 1
  if l = 1
    P1 = Z0
  else
    β = (Rl-1)t Zl-1 / (Rl-2)t Zl-2
    Pl = Zl-1 + β Pl-1
  end
  α = (Rl-1)t Zl-1 / (Pl)t Aj Pl
  Xl = Xl-1 + α Pl
  Rl = Rl-1 - α Aj Pl
end

```

ALGORITHM 2. A multigrid-preconditioned conjugate gradient algorithm.

3.2. Convergence theory. We now outline the convergence theory for the multigrid schemes of the previous section following the approach taken in [8]. The analysis is performed using operator notation and is based upon two conditions. Before introducing the first condition, which concerns the smoothing iteration, it is convenient to define an error operator $K_k = I - R_k A_k$ and its adjoint with respect to the inner product $(A_k \cdot, \cdot)_k$. $K_k^* = I - R_k^t A_k$.

(C.1) There is a constant C_R independent of k such that the smoothing procedure satisfies

$$(3.14) \quad \frac{\|u\|_{0,k}^2}{\lambda_k} \leq C_R (\bar{R}_k u, u)_k \quad \forall u \in M_k,$$

for both $\bar{R}_k = (I - K_k^* K_k) A_k^{-1}$ and $\bar{R}_k = (I - K_k K_k^*) A_k^{-1}$, where λ_k is the largest eigenvalue of A_k , and $\|u\|_{0,k}$ denotes the norm induced by the inner product $(\cdot, \cdot)_k$.

In [7] it is shown that (C.1) is equivalent to the condition that the smoothing iteration (3.8) converge at a rate exceeding that of a Richardson iteration defined by $R_k = \omega \lambda_k^{-1} I$ with $\omega = C_R^{-1}$. That paper then proves (C.1) for a class of smoothers defined by subspace decomposition and which satisfy simple hypotheses. In particular, point, line, and block Jacobi or Gauss-Seidel iterations satisfy (C.1).

REMARK 3. The multigrid algorithms described in [7, 8] use the following iteration as a smoothing

$$x^i = x^{i-1} + R_k^{(i+m(k))} (f_k - A_k x^{i-1}).$$

This iteration differs from (3.8) only notationally. In fact, the meaning given to $R_k^{(1)}$ (e.g. the sweep direction for the first Gauss-Seidel smoothing) can be chosen, perhaps differently, for each notation so that the two iterations are identical. \square

The second condition is expressed in terms of the adjoint $P_{k-1} : M_k \rightarrow M_{k-1}$ of I_k taken with respect to the inner products $A_k(\cdot, \cdot) = (A_k \cdot, \cdot)_k$ and $A_{k-1}(\cdot, \cdot)$,

$$(3.15) \quad A_{k-1}(P_{k-1} u, v) = A_k(u, I_k v), \quad \forall v \in M_{k-1}.$$

(C.2) "Approximation and Regularity" — for some $\alpha \in (0, 1]$ with C_α independent of k ,

$$(3.16) \quad |A_k((I - I_k P_{k-1})u, u)| \leq C_\alpha^2 \left(\frac{\|A_k u\|_{0,k}^2}{\lambda_k} \right)^\alpha A_k(u, u)^{1-\alpha} \quad \forall u \in M_k.$$

The condition (C.2) is typically proved using the approximation properties of the spaces M_k and the elliptic regularity (2.3) of the underlying partial differential equation.

We consider first a result for the W-cycle. For the Morley method, convergence of the W-cycle iteration was first proved by Brenner [9].

THEOREM 4. (W-cycle) *If the conditions (C.1) and (C.2) are satisfied, then the m-smoothing W-cycle iteration defined by Algorithm 1 converges for sufficiently large m, and $E_k = I - B_k A_k$ is a contraction, with contraction number (independent of k) given by*

$$\delta \leq \frac{M}{M + m^\alpha},$$

where M is a constant, $M = M(\alpha, C_\alpha, C_R)$. Furthermore, the same conclusion holds if "m is sufficiently large" is replaced by the assumption

$$(3.17) \quad A_k(I_k u, I_k u) \leq 2 A_{k-1}(u, u) \quad \forall u \in M_{k-1}.$$

An explicit expression for M can be found in [6]. For the bilinear forms $A_k(\cdot, \cdot)$ obtained for the Morley or for the modified Ciarlet-Raviart method, the condition (3.17) fails. And as previously noted, "m sufficiently large" can mean $m \geq 8$ in this context; see Section 4. Not only will the W-cycle diverge for small m , but numerical results show that for the W-cycle, the preconditioned operator $B_j A_j$ may be indefinite. It is not certain that a conjugate gradient iteration for (3.12) will then converge. The next Theorem will demonstrate that one does not encounter this difficulty with the Variable V-cycle.

THEOREM 5. (Variable V-cycle) *If the conditions (C.1) and (C.2) are satisfied, then the Variable V-cycle multigrid algorithm yields preconditioners B_k such that*

$$(3.18) \quad \eta_0 A_k(u, u) \leq A_k(B_k A_k u, u) \leq \eta_1 A_k(u, u), \quad \forall u \in M_k,$$

with

$$(3.19) \quad \eta_0 \geq \frac{m(k)^\alpha}{M + m(k)^\alpha}, \quad \eta_1 \leq \prod_{i=1}^k \left(1 + \frac{C}{m(i)^\alpha}\right) \leq \frac{M + m(k)^\alpha}{m(k)^\alpha},$$

where M is a constant, $M = M(\alpha, C_\alpha, C_R)$.

As consequence of this theorem, the spectral condition number of $B_k A_k$ satisfies

$$\kappa_2(B_k A_k) \leq \frac{\eta_1}{\eta_0} \leq \left(\frac{M + m(k)^\alpha}{m(k)^\alpha} \right)^2,$$

and the multigrid-preconditioned conjugate gradient solution of (3.12) converges with an asymptotic rate of

$$(3.20) \quad \frac{\sqrt{\kappa_2(B_j A_j)} - 1}{\sqrt{\kappa_2(B_j A_j)} + 1} \leq \frac{M}{M + 2m(j)^\alpha}$$

per iteration.

A proof of (C.2) for the Ciarlet-Raviart mixed method may be found in [14]. A proof of (C.2) for the Morley method appears in [5]. The preconditioning properties of the simple V-cycle are questionable in comparison to those of the Variable V-cycle. Numerical experiments described in Section 4 suggest that this condition number may not be bounded independent of the mesh.

4. Numerical Results. According to Theorem 4, convergence of the standard W-cycle iteration applied to the Ciarlet-Raviart or the Morley method is assured *provided* that enough smoothing iterations are performed. Guarantees of convergence are provisional since (3.17) is not satisfied. In this section, we will present computations which show that W-cycle iterations with minimal smoothing are not robust here, these iterations commonly diverge. Rather than abandon multigrid for these finite element methods, we advocate that Theorem 5 be employed in these situations. According to this Theorem, a single Variable V-cycle generates a robust preconditioner for (3.2). The effectiveness of this preconditioner when combined with a preconditioned conjugate gradient (PCCG) iteration is illustrated by a second set of computations.

In the following computations the modification to the Ciarlet-Raviart method and piecewise-quadratic spaces have been used. Using the discrete forms defined by (2.10) makes it affordable to apply Gauss-Seidel smoothing iterations to the Ciarlet-Raviart Schur complement. (This is another good reason to use the modified Ciarlet-Raviart method. Similar multigrid behavior is observed for the standard Ciarlet-Raviart method with Richardson smoothing.) We construct a sequence of nested quasi-uniform triangulations of $\Omega = [0, 1] \times [0, 1]$ from a coarse mesh by joining triangle midpoints to get finer meshes. For each computation, the coarsest mesh consists of two triangles.

The first set of computations examines five-level W-cycles as the number of multigrid smoothing iterations, m , is increased. Eigenvalues λ_{\min} and λ_{\max} of the "W-cycle preconditioned" operators $B_k A_k$ are displayed in Table 1 for $k = 5$ and both Richardson ($\omega = 1.5$) and Gauss-Seidel smoothing. Recall from Section 3.1 that the error reduction operator for the standard multigrid iteration satisfies $E_k = I - B_k A_k$, cf. (3.13). Consequently, if the spectrum of $B_k A_k$ is not contained in the interval $(0, 2)$, then the standard multigrid iteration may diverge. Alternatively, one might consider the multigrid reduction factors $\delta = \max(|\lambda_{\max} - 1|, |1 - \lambda_{\min}|)$. In practice we find that $\delta = \lambda_{\max} - 1$. (This differs from the usual inherited form situation in which $\lambda_{\max} < 1$ and $\delta = 1 - \lambda_{\min}$.) For the modified Ciarlet-Raviart scheme, we see from Table 1 that if Richardson smoothing is used, one must take " $m \geq 8$ " smoothings per five-level W-cycle iteration in order to guarantee convergence. Clearly Gauss-Seidel smoothing is to be preferred for this problem—yet $m > 1$ is still necessary for general W-cycle convergence. Similarly, for the Morley discretization, Richardson smoothing is again unsatisfactory for small m . In this case, it appears that $m = 1$ Gauss-Seidel smoothing yields an acceptable W-cycle. However, it is not certain that this will continue to be true for finer grids or for different domains Ω .

TABLE 1. W-cycle preconditioned $B_5 A_5$, extremal eigenvalues vs. m .

m	modified Ciarlet-Raviart				Morley			
	Richardson		Gauss-Seidel		Richardson		Gauss-Seidel	
	λ_{\max}	λ_{\min}	λ_{\max}	λ_{\min}	λ_{\max}	λ_{\min}	λ_{\max}	λ_{\min}
1	2.48	-23.3	2.23	-1.16	2.10	.211	1.86	.594
2	2.47	-10.5	1.96	.119	1.79	.377	1.65	.815
3	2.29	-1.46	1.84	.520	1.73	.509	1.55	.81
4	2.22	-.645	1.72	.694	1.69	.612	1.49	.922
5	2.18	-.328	1.64	.757	1.65	.694	1.45	.948
6	2.13	-.070	1.56	.806	—	—	—	—
7	2.06	.132	1.49	.844	—	—	—	—
8	1.999	.284	1.44	.874	—	—	—	—

Next we examine the effectiveness of the Variable V-cycle ($\beta = 2$) preconditioners B_k . Computed values for extremal eigenvalues and condition numbers of the preconditioned operators $B_k A_k$, $k = 2, \dots, 6$, are listed in Table 2. According to Theorem 5, the condition number of $B_k A_k$ is bounded independently of k . Further computations with varying β con-

firm that the slow growth of λ_{max} seen in this table is not inconsistent with this Theorem, but is reminiscent of the upper bound in (3.19),

$$\lambda_{max}(B_k A_k) \leq \eta_1 \leq \prod_{i=1}^k \left(1 + \frac{C}{m(i)^\alpha}\right),$$

where $m(i) = 2^{k-i}$. For comparison, note that with $k = 8$ ($h_k = 1/128$), the condition number of A_k , the Schur complement from (2.9), exceeds 10^8 .

From the condition numbers of Table 2 we can estimate reduction factors for Variable V-cycle preconditioned conjugate gradient iterations. For condition numbers $\kappa_2(B_k A_k) \leq 20$ we obtain $\delta_{PCCG} \leq .68$ (averaging over 10 iterations). Similarly, $\delta_{PCCG} \leq .48$ for the Morley method when $\kappa_2(B_k A_k) \leq 7$. These reductions are significantly better than those obtained for W-cycle iterations. According to Table 1, $\delta_{Morley}^{m=1} \leq .86$ and $\delta_{CR}^{m=2} \leq .96$ for $k = 5$. This distinction, which is supported by iteration data (cf. Table 3), is explained in Section 3.2, see (3.9) – (3.12). It was shown there that the standard multigrid iteration is nothing more than a preconditioned iteration using a scheme which is generally less effective than the preconditioned conjugate gradient scheme.

TABLE 2. Extremal eigenvalues for variable V-cycle $B_k A_k$.

k	modified Ciarlet-Raviart			Morley		
	λ_{max}	λ_{min}	$\lambda_{max}/\lambda_{min}$	λ_{max}	λ_{min}	$\lambda_{max}/\lambda_{min}$
2	1.22	0.553	2.20	.9999	0.846	1.18
3	2.18	0.355	6.12	1.47	0.641	2.29
4	3.38	0.297	11.4	2.05	0.612	3.35
5	4.45	0.284	15.7	2.75	0.595	4.63
6	5.07	0.281	18.0	3.40	0.588	5.78

To further illustrate the behavior of the various multigrid schemes we provide Table 3. In particular, this data demonstrates that, in addition to being robust, the Variable V-cycle preconditioner is cheaper to use than an $m > 1$ W-cycle. Recorded in Table 3 are the numbers of iterations required to solve $A_k u_k = f_k$ with relative error less than 2×10^{-7} . Relative error was measured in the norm $(A_k B_k A_k \cdot, \cdot)_k^{1/2}$, (2×10^{-7} corresponds to approximately six correct digits), and iteration counts were averaged for five simple test problems. Results are listed for both the modified Ciarlet-Raviart and the Morley methods. In each case Gauss-Seidel smoothing was used. For the W-cycles, results for the standard multigrid scheme appear first and are followed by a slash and then the results for the associated preconditioned conjugate gradient (PCCG) scheme. Note that the W-cycle may be used as a preconditioner provided that m is large enough. ($m \geq 2$ for the modified Ciarlet-Raviart method). In comparing results note that an ($m = 1$) W-cycle or a ($\beta = 2$) Variable V-cycle iteration is roughly 50 to 100 percent more costly than an ($m = 1$) V-cycle iteration. An m -smoothing W-cycle is roughly m times as expensive as a ($\beta = 2$) Variable V-cycle.

It is not known if the V-cycle provides a bounded preconditioning for these problems.

TABLE 3. Convergence of multigrid iterative and preconditioned schemes.

k	modified Ciarlet-Raviart				Morley		
	variable-V PCCG	W-cycle standard multigrid / PCCG		V-cycle PCCG	variable-V PCCG	W-cycle multigrid / PCCG	V-cycle PCCG
	$\beta = 2$	$m = 2$	$m = 3$	$m = 1$	$\beta = 2$	$m = 1$	$m = 1$
2	6.8	9.0 / 5.0	6.6 / 4.2	6.8	4.8	8.0 / 4.8	4.8
3	15.2	53.8 / 10.8	32.8 / 8.8	14.8	9.0	19.8 / 9.0	8.6
4	20.8	519.0 / 14.0	56.2 / 11.0	21.0	11.6	42.4 / 11.0	11.2
5	24.2	305.6 / 18.8	72.4 / 12.6	26.8	13.4	79.0 / 12.0	14.0
6	26.8	262.6 / 22.0	80.6 / 13.8	32.8	15.0	160.4 / 12.6	16.4
7	27.4	273.4 / 21.8	72.0 / 14.4	41.0	16.4	187.8 / 14.0	19.2
8	26.6	207.6 / 21.0	67.0 / 13.4	49.6	16.6	174.0 / 13.8	22.0

Acknowledgment. The author is indebted to Jim Bramble for many helpful suggestions.

REFERENCES

- [1] D. N. ARNOLD AND F. BREZZI, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Math. Model. Num. Anal., 19 (1985), pp. 7-32.
- [2] R. E. BANK AND C. C. DOUGLAS, *Sharp estimates for multigrid rates of convergence with general smoothing and acceleration*, SIAM J. Numer. Anal., 22 (1985), pp. 617-633.
- [3] D. BRAESS AND W. HACKBUSCH, *A new convergence proof for the multigrid method including the v-cycle*, SIAM J. Numer. Anal., 20 (1983), pp. 967-975.
- [4] D. BRAESS AND P. PEISKER, *On the numerical solution of the biharmonic equation and the role of squaring matrices for preconditioning*, IMA J. Numer. Anal., 6 (1986), pp. 393-404.
- [5] J. H. BRAMBLE, *Multigrid methods*. Cornell University Mathematics Department Lecture Notes, 1992.
- [6] J. H. BRAMBLE AND J. E. PASCIAK, *New convergence estimates for multigrid algorithms*, Math. Comp., 49 (1987), pp. 311-329.
- [7] ———, *The analysis of smoothers for multigrid algorithms*, Math. Comp., 58 (1992), pp. 467-488.
- [8] J. H. BRAMBLE, J. E. PASCIAK, AND J. XU, *The analysis of multigrid algorithms with non-embedded spaces or non-inherited quadratic forms*, Math. Comp., 56 (1991), pp. 1-34.
- [9] S. C. BRENNER, *An optimal-order nonconforming multigrid method for the biharmonic equation*, SIAM J. Numer. Anal., 26 (1989), pp. 1124-1138.
- [10] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, vol. 4 of Stud. Math. Its Appl., North-Holland, Amsterdam, 1978.
- [11] P. G. CIARLET AND P. RAVIART, *A mixed finite element method for the biharmonic equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, New York, 1974, pp. 125-145.
- [12] G. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, second ed., 1989.
- [13] P. GRISVARD, *Elliptic Problems in Non-smooth Domains*, vol. 24 of Monographs and Studies in Mathematics, Pitman, Boston, 1985.
- [14] M. R. HANISCH, *Multigrid preconditioning for the biharmonic Dirichlet problem*, to appear in SIAM J. Numer. Anal.
- [15] ———, *Multigrid Preconditioning for Mixed Finite Element Methods*, PhD thesis, Cornell University, Ithaca, NY, 1991.
- [16] L. HERRMANN, *A bending analysis for plates*, in Proc. Conference on Matrix Methods in Structural Mechanics, Dayton, Ohio, 1966, Air Force Flight Dynamics Laboratory, pp. 577-604. AFFDL-TR-66-80.
- [17] ———, *Finite element bending analysis for plates*, J. Engrg. Mech. Div. A.S.C.E. EM5, 93 (1967), pp. 13-26.
- [18] C. JOHNSON, *On the convergence of a mixed finite element method for plate bending problems*, Numer. Math., 21 (1973), pp. 43-62.
- [19] J. MANDEL, *Algebraic study of multigrid methods for symmetric, definite problems*, Appl. Math. Comput., 25 (1988), pp. 39-56.
- [20] J. MANDEL, S. MCCORMICK, AND R. BANK, *Variational multigrid theory*, in Multigrid Methods, S. McCormick, ed., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1987, pp. 131-178.
- [21] J. MANDEL, S. MCCORMICK, AND J. RUGE, *An algebraic theory for multigrid methods for variational problems*, SIAM J. Numer. Anal., 25 (1988), pp. 91-110.
- [22] T. MIYOSHI, *A finite element method for the solution of fourth order partial differential equations*, Kumamoto J. Sci. (Math.), (1973), pp. 87-116.
- [23] L. S. D. MORLEY, *The triangular equilibrium element in the solution of plate bending problems*, Aero. Quart., 15 (1968), pp. 149-169.
- [24] P. PEISKER, *A multilevel algorithm for the biharmonic problem*, Numer. Math., 46 (1985), pp. 623-634.
- [25] P. RAVIART AND J. THOMAS, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, Springer-Verlag, Berlin, 1977, pp. 292-315.

SIMULATED ANNEALING ALGORITHMS FOR CONTINUOUS OPTIMIZATION

Saul B. Gelfand, Peter C. Doerschuk and Mohamed Nahhas-Mohandes
School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907-1285

ABSTRACT. Simulated annealing algorithms for optimization over continuous spaces come in two varieties: Markov chain algorithms and modified gradient algorithms. Unfortunately, there is a gap between the theory and the application of these algorithms: the convergence conditions cannot be practically implemented. In this paper we suggest a practical methodology for implementing the modified gradient annealing algorithms based on their relationship to the Markov chain algorithms.

1. INTRODUCTION. Simulated annealing is a popular approach to global optimization of functions with multiple local minima. One type of annealing algorithm for continuous optimization involves simulating a Markov chain using a generalized Metropolis (or related) method. We refer to these algorithms as Markov chain annealing algorithms (MCAA's). There is a large amount of theoretical analysis and practical methodology developed for the MCAA's (Vanderbilt and Louie, 1984; Bohachevsky et al., 1986; Corana et al., 1987; Brooks and Verdini, 1988; Press and Teukolsky, 1991; Gelfand and Mitter, 1992). However, the feasibility of MCAA's for high-dimensional problems is questionable.

Another type of annealing algorithm for continuous optimization involves modifying gradient-type search algorithms. Let $U(\bullet)$ be a smooth cost function on \mathbb{R}^D . A standard gradient algorithm for finding a local minimum of $U(\bullet)$ (and hence a global minimum if $U(\bullet)$ is convex) is given by

$$z_{k+1} = z_k - \mu \nabla U(z_k)$$

where μ is a step-size parameter. A modified gradient algorithm for finding a global (or near global) minimum of $U(\bullet)$ is given by

$$X_{k+1} = X_k - \mu \nabla U(X_k) + \sqrt{2T\mu} W_k$$

where $\{W_k\}$ is a white Gaussian noise sequence and T is a "temperature" parameter which is slowly decreased as the algorithm proceeds. The idea behind this algorithm is that by artificially adding in the noise term (via Monte Carlo simulation) it is possible to escape from strictly local minima. We refer to this modified gradient algorithm as a gradient annealing algorithm (GAA). Now there is some theoretical analysis developed for the gradient annealing algorithm (Kushner, 1987; Gelfand and Mitter, 1991b,c), but no practical methodology that we are aware of. On the other hand, there may be some hope of using GAA for high-dimensional problems with smooth well-behaved cost functions, as it attempts to exploit the smoothness by its use of derivatives. The goal of this paper is to use some theory from Gelfand and Mitter (1991a) relating the MCAA and GAA, and some practical methodology from Johnson et al. (1989) for the MCAA, to develop a practical methodology for GAA.

2. MARKOV CHAIN ANNEALING ALGORITHMS. Most of the theory and application of MCAA deals with discrete (combinatorial) optimization. The literature on MCAA's for continuous optimization is by and large a straightforward generalization of the discrete case. It is this point of view we discuss in this section. The discussion is very brief and the reader is referred to the literature for more details.

Let $U(\bullet)$ be a cost function on \mathbb{R}^D . We wish to find an element of \mathbb{R}^D which minimizes $U(\bullet)$. A general description of the MCAA for solving this problem is as follows (we only consider the Metropolis procedure here):

Given a current solution $x \in \mathbb{R}^D$ generate a candidate solution $y \in \mathbb{R}^D$

If $U(y) \leq U(x)$ then accept y as the next solution.

If $U(y) > U(x)$ then accept y as the next solution with probability $\exp(-(U(y) - U(x))/T)$; (otherwise the next solution is the current solution x).

Here the candidate solution is usually a probabilistically generated perturbation of the current solution. Also, the "temperature" parameter T is slowly decreased as the algorithm proceeds, making transitions to higher cost states less likely. The algorithm stops subject to some termination criterion.

The MCAA can be precisely formulated as a continuous state Markov chain as follows. Let $q(x,y)$ be a transition probability density from x to y ($x, y \in \mathbb{R}^D$); $q(x,y)$ is a probability density for the candidate state y given the current state x . The continuous state annealing chain $\{Y_k\}$ (at a fixed temperature T) has 1-step transition probability density from x to y given by

$$(2.1) \quad p(T, x, y) = s(T, x, y)q(x, y) + m(T, x)\delta(y - x)$$

where

$$s(T, x, y) = \exp \left[- \frac{[U(y) - U(x)]^+}{T} \right]$$

and $m(T, x)$ is chosen to provide the correct normalization. Here $[\bullet]^+$ denotes positive part and $\delta(\bullet)$ is a Dirac-delta function. For a fixed temperature T this annealing chain $\{Y_k\}$ has a Gibbs equilibrium distribution with density function

$$\pi(T, x) = \frac{1}{Z(T)} \exp \left[- \frac{U(x)}{T} \right];$$

$$Z(T) = \int \exp \left[- \frac{U(x)}{T} \right] dx (< \infty),$$

and as the temperature T tends to zero we get $\pi(T, \bullet)$ converging to a density $\pi^*(\bullet)$, which is concentrated on the global minima of $U(\bullet)$. If the rate of temperature decrease is slow enough, then $\{Y_k\}$ remains near the equilibrium distributions and also concentrates on the global minima for k large (Gelfand and Mitter, 1992) (we note that the proof of convergence in the continuous case naturally requires many more technical assumptions and details than

the discrete case).

Unfortunately, there is a large gap between the theory and application of the MCAA. The main problem is that the theoretically appropriate rates of decrease for the temperature are far too slow for practical implementation. In practice, one needs a temperature schedule, a candidate generator and a termination criterion which achieve desirable tradeoffs between complexity and performance.

A practical methodology for continuous state MCAA's can be adopted with relatively few changes from the methodology for discrete state MCAA's developed by Johnson et al. (1989) (refinements of this latter methodology form the basis for most implementations of MCAA's in both continuous and discrete state-space). A key quantity in this methodology is the acceptance probability $P_A(T)$ which is estimated by

$$(2.2) \quad \hat{P}_A(T) = \frac{N_A(T)}{N(T)}$$

where $N_A(T)$ is the number of moves accepted, and $N(T)$ is the number of moves attempted, at temperature T . Although there is some motivation for allowing $N(T)$ to increase with decreasing T , the experiments by Johnson et al. (1989) suggest that there is no real advantage to doing so, and hence $N(T)$ is fixed at some number N . The methodology proceeds by making this fixed number of iterations (attempted moves) at each of a sequence of geometrically decaying temperatures, and the initial and final temperatures are selected by requiring that the acceptance probability be specified values. For termination in the discrete case it is also required that the running cost for the best solution has not decreased over the 5 previous temperature values; in the continuous case considered here we modify this to only require that the running cost for the best solution has not decreased by more than a small threshold. A summary description of the algorithm is given below.

Markov chain annealing algorithm methodology. Input parameters: p_0 (initial acceptance probability), p_F (final acceptance probability), ρ (geometric ratio in temperature schedule), N (number of iterations at any temperature), ϵ (termination threshold)

1. Find initial temperature T_0 such that $\hat{P}_A(T_0) = p_0$, ($\hat{P}_A(T_0)$ given by Equation (2.2)).
2. Set $j = 0$.
3. Run the annealing chain N iterations at temperature $T_j = \rho^j T_0$
4. Let Y_j^* be the best solution found through temperature T_j
 If $\hat{P}_A(T_j) \leq p_F$ and $U(Y_j^*) \geq U(Y_{j-5}^*) - \epsilon$
 then terminate the search and output Y_j^* and $U(Y_j^*)$
 else set $j = j + 1$ and go to 3. □

Although there have been some successes reported with MCAA's of the general type described above, it has been observed that the method is very inefficient for high dimensional problems, essentially because it does not exploit the smoothness of the cost function. In the next section, we discuss the GAA which may overcome this inefficiency in some problems.

3. GRADIENT ANNEALING ALGORITHM. Let $U(\cdot)$ be a smooth cost function (at least C^2) on \mathbb{R}^D . We wish to find an element of \mathbb{R}^D which minimizes $U(\cdot)$. Here we consider GAA as an alternative to the MCAA described in Section 2.

The GAA (with a fixed step size μ and temperature T) is given by the following stochastic recursion

$$(3.1) \quad X_{k+1} = X_k - \mu \nabla U(X_k) + \sqrt{2T\mu} W_k$$

where $\{W_k\}$ is a standard D -dimensional white Gaussian noise sequence, artificially added in (via Monte Carlo simulation) to try to avoid getting trapped in local minima, and the temperature T (and possibly the step-size μ) is slowly decreased as k gets large. The asymptotic (large-time) behavior of GAA and MCAA are similar. For fixed temperature T and small step-size μ the process $\{X_k\}$ *nearly* has a Gibbs equilibrium distribution with density function $\pi(T, \cdot)$, and as the temperature T tends to zero we get $\pi(T, \cdot)$ converging to a $\pi^*(\cdot)$ which is concentrated on the global minima of $U(\cdot)$. If the rate of temperature and step-size decrease are chosen appropriately, then $\{X_k\}$ remains near the equilibrium distributions and also concentrates on the global minima for large k (Kushner, 1987; Gelfand and Mitter, 1991b,c).

GAA is plagued by the same gap between theory and application as the MCAA's. Theoretically appropriate rates of decrease for the temperature schedule are too slow for practical implementation, and no results are available concerning the important case of fixed step-size which by analogy with standard gradient algorithms is necessary for rapid convergence. In practice one needs a temperature schedule, a step-size (assumed known and fixed here) and a termination criterion which achieve desirable tradeoffs between complexity and performance. Practical implementation of GAA appears not to have received any attention in the literature.

We shall suggest a practical methodology for GAA based on the methodology for MCAA's discussed in Section 2, and the relationship between GAA and MCAA which we shall elaborate on below. We shall show that GAA and a certain class of MCAA interpolated into continuous time (with step-size/interpolation interval μ) both have a diffusion limit (as $\mu \rightarrow 0$), and these diffusion limits are linearly time-scaled versions of one another. Hence by taking into account the appropriate time-scaling, we can use the MCAA methodology as a basis for a GAA methodology. An important feature of this approach is that it allows us to implicitly associate the idea of acceptance probability with GAA - a critical quantity in developing temperature schedules and initialization and termination criterion for most practical annealing schemes.

3.1. Diffusion Limits for MCAA and GAA. We first formulate a MCAA which has the appropriate structure and scaling to admit a diffusion limit. Referring to the general version of the MCAA in Section 2 we consider here a transition density $q(\cdot, \cdot)$ which corresponds to selecting a coordinate direction at random, and then making a Gaussian perturbation along that coordinate. Let x_i denote the i -th coordinate of $x \in \mathbb{R}^D$. We choose

$$(3.2) \quad q(x, y) = \frac{1}{D} \sum_{i=1}^D N(x_i, \alpha\mu)(y_i) \prod_{j \neq i} \delta(y_j - x_j)$$

where $N(m, \sigma^2)(\bullet)$ denotes a (scalar) Gaussian density with mean m and variance σ^2 . Note that the variance of the Gaussian perturbation along the selected coordinate is $\alpha\mu$ where α does not depend on T (so that $q(\bullet, \bullet)$ does not depend on T) and is to be specified. Let $\{Y_k\}$ be a Markov chain with 1-step transition density $p(T, \bullet, \bullet)$ given by Equations (2.1) and (3.2). Interpolate $\{Y_k = Y_k^\mu: k = 0, 1, \dots\}$ into a continuous-time process $\{Y^\mu(t): t \geq 0\}$ by

$$Y^\mu(t) = Y_k^\mu, \quad t \in [k\mu, (k+1)\mu), \quad k = 0, 1, \dots$$

It can be shown (Gelfand and Mitter, 1991a) that $Y^\mu(\bullet)$ has a diffusion limit, i.e., $Y^\mu(\bullet) \rightarrow Y(\bullet)$ as $\mu \downarrow 0$ (in law), where $Y(\bullet)$ satisfies the Ito equation

$$dY(t) = -\frac{\alpha}{2TD} \nabla U(Y(t))dt + \sqrt{\frac{\alpha}{D}} dV(t)$$

and $V(\bullet)$ is a standard D -dimensional Wiener process.

Next, consider the GAA $\{X_k\}$ given by Equation (3.1). Interpolate $\{X_k = X_k^\mu: k = 0, 1, \dots\}$ into a continuous-time process $\{X^\mu(t): t \geq 0\}$ by

$$X^\mu(t) = X_k^\mu, \quad t \in [k\mu, (k+1)\mu), \quad k = 0, 1, \dots$$

It is easy to show that $X^\mu(\bullet)$ also has a diffusion limit, i.e., $X^\mu(\bullet) \rightarrow X(\bullet)$ as $\mu \downarrow 0$ (in law), where $X(\bullet)$ satisfies the Ito equation

$$dX(t) = -\nabla U(X(t))dt + \sqrt{2T} dW(t)$$

and $W(\bullet)$ is a standard D -dimensional Wiener process.

Now consider the process defined by linearly scaling time by a factor $\beta > 0$ in the process $X(\bullet)$:

$$\tilde{X}(t) = X(\beta t)$$

By standard calculations $\tilde{X}(\bullet)$ satisfies the Ito equation

$$d\tilde{X}(t) = -\beta \nabla U(\tilde{X}(t))dt + \sqrt{2T\beta} d\tilde{W}(t)$$

where $\tilde{W}(\bullet)$ is a standard D -dimensional Wiener process. From the (assumed) uniqueness of the Ito equation solution, it is seen that if we take

$$\beta = \beta(T) = \frac{\alpha}{2TD}$$

then $Y(\bullet) = \tilde{X}(\bullet)$ (in law), i.e., $Y(\bullet)$ is a linearly time-scaled version of $X(\bullet)$, with scale-factor $\beta(T)$ depending inversely on the temperature T .

3.2. Toward a Methodology for GAA. In view of the limit diffusion behavior exhibited by both GAA and MCAA, we have that under suitable conditions, GAA is close to a linearly

time-scaled version of MCAA. This suggests that we can use the MCAA methodology to guide the GAA methodology by correcting for the scaling (this is not to say that GAA and MCAA perform the same; see the discussion in Section 4).

The idea is the following. Suppose we run the MCAA $\{Y_k\}$ for N iterations at temperature T , and consider the GAA $\{X_k\}$ also at temperature T . Then since the limit diffusion $Y(\bullet)$ for $\{Y_k\}$ is a linearly time-scaled version of the limit diffusion $X(\bullet)$ for $\{X_k\}$ with scale factor $\beta(T)$, the suggestion is to run the GAA $\{X_k\}$ for $N(T) = \beta(T)N$ iterations at temperature T to compensate for the time scaling. Now we still need to choose the parameter α in the variance of the MCAA. We do this by choosing $\beta(T_0) = 1$, i.e., we choose the GAA and MCAA to run at the same time scale at the initial (high) temperature value T_0 (this choice is somewhat arbitrary but avoids introducing additional parameters). Hence $\alpha = 2T_0D$ and so $\beta(T) = T_0/T$ and thus $N(T) = (T_0/T)N$, and we run the GAA $\{X_k\}$ for

$$N_j = \rho^{-j} N_0$$

iterations at temperature $T_j = \rho^j T_0$. From a practical point of view, we may impose a ceiling on the number of iterations the GAA can make at any temperature.

The basic structure of the MCAA methodology now carries over to a GAA, except that the MCAA uses a fixed number of iterations at each of a geometrically decreasing sequence of temperatures, while the GAA uses a geometrically increasing sequence of iterations at a geometrically decreasing sequence of temperatures.

To apply the MCAA methodology to GAA it is desirable to find a good estimate of the acceptance probability, which is used to determine the initial and final temperatures. Clearly, it is not desirable to estimate the acceptance probability via a Monte Carlo simulation of a MCAA (in addition to the GAA). Now in view of the limit diffusion analysis, the appearance of the gradient term in the GAA can be viewed as a local approximation in a certain MCAA. This approximation is possible because of the (assumed) smoothness in the cost function and the smallness of the step size, and should result in significant computational and performance advantages for GAA. We shall next discuss how to make some other local approximations in the MCAA to facilitate estimation of the acceptance probability, which should make for more efficient determination of the initial and final temperatures.

The acceptance probability at temperature T is given by

$$P_A(T) = \int \pi(T, x) P_A(T | x) dx$$

where

$$P_A(T | x) = \int s(T, x, y) q(x, y) dy$$

is the conditional probability of accepting a candidate move given the current state is x . We develop an approximation to $P_A(T)$ as follows. Substituting for $q(\bullet, \bullet)$ from Equation (3.2) and setting $\alpha = 2T_0D$ we can write

$$P_A(T|x) = \frac{1}{D} \sum_{i=1}^D P_A(T|x,i)$$

where

$$P_A(T|x,i) = \int s(T,x,y) N(x_i, 2T_0 D \mu)(y_i) \prod_{j \neq i} \delta(y_j - x_j) dy$$

is the conditional probability of accepting a candidate move given the current state is x and coordinate i is selected for perturbation. Fix T , x and i for the moment and let

$$\tilde{x} = (x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_D)$$

Then

$$P_A(T|x,i) = \int \exp \left[-\frac{[U(\tilde{x}) - U(x)]^+}{T} \right] N(x_i, 2T_0 D \mu)(y_i) dy_i$$

We estimate $P_A(T|x,i)$ by

$$\hat{P}_A(T|x,i) = \int \exp \left[-\frac{[U_{x_i}(x)(y_i - x_i) + U_{x_i x_i}(x)(y_i - x_i)^2/2]^+}{T} \right] \cdot N(x_i, 2T_0 D \mu)(y_i) dy_i$$

It is possible to work out expressions for $\hat{P}_A(T|x,i)$ in terms of the exponential and error functions. Finally we estimate $P_A(T)$ by

$$(3.3) \quad \hat{P}_A(T) = \frac{1}{N(T)} \sum_{k=1}^{N(T)} \hat{P}_A(T|X_k)$$

$$\hat{P}_A(T|x) = \frac{1}{D} \sum_{i=1}^D \hat{P}_A(T|x,i)$$

where the average in the first equation is computed over the $N(T)$ iterations of GAA at temperature T .

A summary description of the proposed methodology for the GAA is given below.

Gradient Annealing Algorithm Methodology Input parameters: p_0 (initial acceptance probability), p_F (final acceptance probability), ρ (geometric ratio in temperature schedule), N_0 (number of iterations at initial temperature), ϵ (termination threshold)

1. Find initial temperature T_0 such that $\hat{P}_A(T_0) = p_0$ ($\hat{P}_A(T_0)$ given by Equation (3.3)).
2. Set $j = 0$.

3. Run the modified gradient algorithm $N_j = \rho^{-j} N_0$ iterations at temperature $T_j = \rho^j T_0$
4. Let X_j^* be the best solution found through temperature T_j
 If $\hat{P}_A(T_j) \leq p_f$ and $U(X_j^*) \geq U(X_{j-5}^*) - \varepsilon$
 then terminate the search and output X_j^* and $U(X_j^*)$
 else set $j = j+1$ and go to 3. □

4. CONCLUSIONS. In this paper we have developed a methodology for GAA based on the relationship between GAA and MCAA. The idea here is that GAA and a certain MCAA have diffusion limits which are linearly time-scaled versions of each other, which suggests that a GAA methodology can be obtained from a MCAA methodology by correcting for the time-scaling. This approach allows us to associate the idea of acceptance probability with GAA, a quantity which plays a critical role in temperature schedules for most practical annealing schemes. We also show how to make some local approximations which facilitate better estimation of the acceptance probability.

The experimental evaluation of the proposed GAA methodology is currently being undertaken. One interesting comparison would be between GAA and MCAA. Our intuition is that GAA will do a better job of finding a global minimum than MCAA for sufficiently smooth, well-behaved cost functions. For this to make sense, the implicit assumption that we are making is that GAA and MCAA are *close enough* to their diffusion limit to have a similar methodology (i.e., structure of their temperature schedule and termination criterion), but *far enough* from their diffusion limit to have distinctly different performance on certain problems.

REFERENCES

- Bohachevsky, I.O., M.E. Johnson, and M.L. Stein. 1986. Generalized simulated annealing for function optimization. *Technometrics* 28:209-217.
- Brooks, D.G. and W.A. Verdin. 1988. Computational experience with generalized simulated annealing over continuous variables. *American Journal Mathematics and Management Sciences* 8:425-449.
- Corana, A., M. Marchesi, C. Martini, and S. Ridella. 1987. Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. *ACM Transactions on Mathematical Software* 13:262-280.
- Gelfand, S.B. and S.K. Mitter. 1991a. Weak convergence of Markov chain sampling methods and annealing algorithms to diffusions. *Journal of Optimization Theory and Applications* 68:483-498.
- Gelfand, S.B. and S.K. Mitter. 1991b. Simulated annealing type algorithms for multivariate optimization. *Algorithmica* 6:419-436.
- Gelfand, S.B. and S.K. Mitter. 1991c. Recursive stochastic algorithms for global optimization in \mathbb{R}^d . *SIAM Journal on Control and Optimization* 29:999-1018.
- Gelfand, S.B. and S.K. Mitter. 1992. Metropolis-type annealing algorithms for global optimization in \mathbb{R}^d . *SIAM Journal on Control and Optimization* (to appear).
- Johnson, D.S., C.R. Aragon, L.Y. McGeoch and C. Schevon. 1989. Optimization by simulated annealing: an experimental evaluation: Part I, graph partitioning. *Operations Research* 37:865-892.
- Kushner, H.J.. 1987. Asymptotic global behavior for stochastic approximation and

- diffusions with slowly decreasing noise effects: global minimization via Monte Carlo. *SIAM Journal on Applied Mathematics* 47:169-185.
- Press, W.H. and S.A. Teukolsky. 1991. Simulated annealing optimization over continuous spaces. *Computers in Physics* July/Aug:426-429.
- Vanderbilt, D. and S.G. Louie. 1984. A Monte Carlo simulated annealing approach to optimization over continuous variables. *Journal of Computational Physics* 56:259-271.

RESPONSE OF A CYLINDRICAL SECTION TO AN EXPLOSIVE BLAST

Aaron Das Gupta
Research Mechanical Engineer
U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005-5066.

ABSTRACT

Transient response analysis of a hollow long cylindrical section subjected to an internal explosive blast has been conducted from a structural integrity standpoint. The reflected blast overpressure upon the internal surface of the cylinder was estimated based on a cube root scaling law and the modified Friedlander exponential decay. A closed form solution of the governing equation of motion for the internally pressurized cylinder was obtained subject to initial conditions. The solution was optimized by a trial and error approach to yield an optimum time which was employed to predict the peak response.

INTRODUCTION

There is a continuing need for modeling the structural damage due to explosive blast to ensure structural integrity and to rationally provide hardening for such structures. This has been a subject of long standing interest for the U.S. Army and the Ballistic Research Laboratory (BRL). A number of studies have been performed and damage data [1-6] gathered over the past several years. However, most data available at present are in the form of impulse correlation curves and residual deformation and relatively little has been devoted to transient response studies due to explosive blast effects on structures at high strain rates.

Recently, computation using hydrodynamic codes for shock wave propagation and structural loading estimation [7-9] has been reported. Unfortunately the application of such codes to generate loading and subsequent coupling with structural response prediction codes is rather expensive and laborious. Such procedures are justifiable when a high degree of accuracy is needed in modeling complex problems involving large deformation and nonlinearity. A detailed analysis with coupled codes is beyond the scope of the current investigation due to time and cost constraints. This investigation involves the development of a simplified model and formulation of the equation of motion as well as subsequent effort at a closed-form solution based on simplifying assumptions and previous work on blast loaded plates [10-19] to obtain optimum structural response and critical time of occurrence for the structure.

ESTIMATION OF TRANSIENT LOADS

The transient loads were estimated under the assumption that the detonation of the explosive would generate a blast wave which would impose an uniform reflected overpressure on the internal surface of a long cylindrical hollow tube section with open ends. Although some variation in internal pressure peaks, arrival and duration times at different locations of the section can be expected due to variation in scaled distance from the point of detonation to the surface of the cylinder, for the sake of simplicity, this variation has been ignored in the current investigation.

For the estimation of reflected overpressure loading, a conservative cube-root scaling law [18,19] is employed to compute the scaled distance, z , from the charge location to the nearest location of the wall in the form

$$z = R/W_E^{1/3} \quad (1)$$

where W_E is the equivalent charge weight and R is the distance of the nearest point of the wall from the charge location.

Once the scaled distance is known, the reflected parameters such as peak overpressure, impulse, duration time and time of arrival of the blast wave could be estimated from compiled air blast tables or the Blast code [20-23]. The decay of the reflected overpressure is assumed to obey the modified Friedlander exponential decay equation which can be written as

$$P(t) = P_m[1 - t/t_p]e^{-a't/t_p} \quad (2)$$

where t_p is the positive pressure phase duration of the impulse measured from the time of arrival of the blast wave front at the nearest internal cylindrical wall location, P_m is the peak reflected pressure in excess of the ambient condition and t is the total elapsed time.

The total reflected impulse imposed on the surface of the cylinder can be obtained by integrating the pressure with respect to time from initial up to the positive phase duration as shown below

$$I_r = \int_0^{t_p} P(t)dt \quad (3)$$

where we have tacitly assumed that time zero is the initial time when the blast wave reaches the internal wall such that arrival times could be ignored. Upon substituting the expression for $P(t)$ from equation (2) in above and performing the integration as indicated earlier results in

$$I_r = P_m t_p [a' - 1 + e^{-a'}]/(a')^2 \quad (4)$$

where I_r is the reflected impulse. Using the value of z from equation (1), one can obtain the values for I_r , P_m and t_p from Reference [21]. The value of a' which is the decaying exponent can be determined using the above three quantities in equation (4) by a trial and error implicit solution. When a' is determined, the complete reflected pressure-time loading history upon the structure is known.

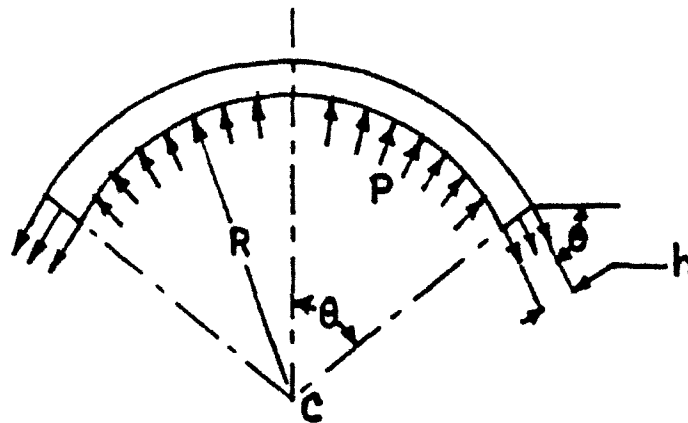


Fig.1. Free body diagram of a pressurized cylindrical section.

FORMULATION OF THE EQUATION OF MOTION

The equation of motion for a hollow cylindrical section subjected to an internal blast from the detonation of a symmetrically located charge weight can be obtained by considering the free-body diagram in Figure 1.

Imposing a balance of forces in the radial direction yields the governing relationship for the cylindrical section as follows:

$$P(t)R\theta - \sigma_{\theta\theta}h\sin(\theta) = M\partial^2 u_r / \partial t^2 \quad (5)$$

where $P(t)$ is the internal transient pressure, R is the internal radius M is the mass per unit length given by $M = \rho R \theta h$, ρ is the mass density of the material from which the cylinder has been fabricated, $\partial^2 u_r / \partial t^2$ is the acceleration of the differential section in the radial direction, $\sigma_{\theta\theta}$ is the circumferential stress in the hoop direction and the remaining variables are defined in Figure 1. To obtain the transient structural response, the differential equation above needs to be reduced to a form amenable to a closed-form solution subject to appropriate initial and boundary conditions.

For a differential circumferential segment as $\sin(\theta)$ approaches θ which can be replaced by $d\theta$ in the differential form, equation (5) could be simplified in the form:

$$\partial^2 u_r / \partial t^2 + \sigma_{\theta\theta} / (\rho R) = P(t) / (\rho h) \quad (6)$$

In order to solve the above equation, we need to obtain expressions for $\sigma_{\theta\theta}$ and $P(t)$ as a function of u_r .

The strain-displacement relations for the normal components of a long hollow cylinder with axisymmetric loading are given as

$$\begin{aligned}\epsilon_{rr} &= \partial u_r / \partial R, \\ \epsilon_{\theta\theta} &= u_r / R, \\ \text{and, } \epsilon_{zz} &= 0.\end{aligned}\tag{7}$$

Assuming linear elastic conditions and noting that volumetric stress in the radial direction is rather small for the long hollow cylinder, an expression for volumetric strain in the hoop direction can be derived as

$$\epsilon_{\theta\theta} = -[(1 - \nu)/\nu]\epsilon_{rr}\tag{8}$$

Invoking Hooke's Law, circumferential stress in the cylinder could be obtained as

$$\sigma_{\theta\theta} = [E/(1 + \nu)(1 - 2\nu)][\nu\epsilon_{rr} + (1 - \nu)\epsilon_{\theta\theta}]\tag{9}$$

where ϵ_{rr} and $\epsilon_{\theta\theta}$ are volumetric strain in the radial and circumferential directions respectively, ν is Poisson's ratio and E is the modulus of elasticity of the material. Substituting ϵ_{rr} and $\epsilon_{\theta\theta}$ from equations (7,8) in above results in an expression for the circumferential hoop stress as a function of the radial displacement and the internal averaged radius which could be given as

$$\sigma_{\theta\theta} = Eu_r/[R(1 - \nu^2)]\tag{10}$$

Substitution of equation (2) and above in the equation of motion developed in the previous section results in a differential equation of motion in the form

$$\partial u_r^2 / \partial t^2 + \omega^2 u_r = (\alpha - \beta t)e^{-\gamma t}\tag{11}$$

where,

$$\begin{aligned}\omega^2 &= E/[\rho R^2(1 - \nu^2)] \\ \alpha &= P_m/(\rho h) \\ \beta &= P_m/(\rho h t_p) \\ \gamma &= a'/t_p\end{aligned}\tag{12}$$

METHOD OF SOLUTION

The governing equation of motion for the cylinder wall must be solved subject to initial conditions that both initial displacement and velocity are zero at time $t = 0$. Let us assume a trial solution for radial displacement of the form

$$u_r(t) = A \sin \omega t + B \cos \omega t + (C - Dt)e^{-\gamma t}\tag{13}$$

Substituting the trial solution above in equation (11) and setting the coefficient of the functions $e^{-\gamma t}$ and $te^{-\gamma t}$ to zero, the constants C and D can be evaluated to obtain

$$u_r(t) = A \sin \omega t + B \cos \omega t + (1/\zeta)[\alpha - 2\gamma\beta/\rho - \beta t]e^{-\gamma t} \quad (14)$$

where,

$$\zeta = \gamma^2 + \omega^2 \quad (15)$$

Subsequently, the initial conditions are applied in the equation above to evaluate the constants A and B which yields

$$\zeta u_r(t) = [\delta + \beta/\gamma](\gamma/\omega) \sin \omega t - \delta \cos \omega t + (\delta - \beta t)e^{-\gamma t} \quad (16)$$

where,

$$\delta = \alpha - 2\gamma\beta/\zeta \quad (17)$$

The above equation computes in a closed-form the time dependent radial displacement variation of the cylinder as a function of the geometric and material parameters as well as the transient internal pressure load due to detonation of an explosive.

To obtain peak radial displacement of the cylinder, the time of occurrence of the peak radial response must be computed. Taking derivative of the radial displacement with respect to time and setting it to zero ensures that the displacement attains an optimum. This operation ensures that the radial displacement is an optimum at an elapsed time provided the radial velocity is zero. However, to guarantee that the displacement is in fact a maximum, the radial acceleration should also be negative. The radial acceleration can be obtained by differentiating the velocity again with respect to time which should vanish when radial velocity attains a peak value.

DETERMINATION OF PEAK STRESS

Using the definitions of $\sigma_{\theta\theta}$ from equation (10) and ω^2 from (12), it can be easily seen that

$$\sigma_{\theta\theta} = \rho R \omega^2 u_r \quad (18)$$

Substituting u_r from equation (19) in above yields

$$\sigma_{\theta\theta} = \rho R \omega^2 / \zeta [(\delta + \beta/\gamma)(\gamma/\omega) \sin \omega t - \delta \cos \omega t + (\delta - \beta t)e^{-\gamma t}] \quad (19)$$

The above equation can be used to determine the maximum hoop stress that the cylindrical wall will experience in case of an accidental explosion. However, prior to peak stress computation, the time of occurrence of the peak response must be found. Evaluating the derivative of $\sigma_{\theta\theta}$ with respect to time at the optimum time of occurrence and setting the resulting expression to zero, we obtain

$$(\delta\gamma + \beta)\cos(\omega t_m) + \delta\omega\sin(\omega t_m) - [\gamma(\delta - \beta t_m) + \beta]e^{-\gamma t_m} = 0 \quad (20)$$

Above equation can be solved implicitly for the optimum time, t_m , when hoop stress becomes a maximum. To ensure that the stress is indeed a maximum, the left hand side expression in above must be differentiated again with respect to time and evaluated to ensure that the double derivative of $\sigma_{\theta\theta}$ with respect to time is negative at the optimum time.

Once the value of the optimum time, t_m is known, it can be substituted in equation (19) to obtain the maximum hoop stress in the form

$$\sigma_{peak} = \rho R \omega^2 / \zeta [(\delta + \beta/\gamma)(\gamma/\omega)\sin(\omega t_m) - \delta\cos(\omega t_m) + (\delta - \beta t_m)e^{-\gamma t_m}] \quad (21)$$

Peak radial displacement can be predicted by substituting the optimum time of occurrence in equation (17) which yields

$$(u_r)_{peak} = (1/\zeta)[(\delta + \beta/\gamma)(\gamma/\omega)\sin(\omega t_m) - \delta\cos(\omega t_m) + (\delta - \beta t_m)e^{-\gamma t_m}] \quad (22)$$

Once peak radial displacement is known, peak radial velocity and acceleration could be determined by taking single and double derivatives of the displacement given in equation (17) and replacing t by t_m in the corresponding terms.

REFERENCES

1. Hanna, J.W., "An Effectiveness Evaluation of Several Types of Antitank Mines," BRL-MR-616, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, June 1952. (AD 377342)
2. Norman, R.M., "Deformation in Flat Plates Exposed to HE Mine Blast," AMSAA-TM-74, US Army Materiel Systems Analysis Agency, Aberdeen Proving Ground, MD, May 1970.
3. Hoskin, N.E., Allan, J.W., Bailey, W.A., Lethaby, J.W. and Skidmore, I., "The motion of plates and cylinders driven at tangential incidence", Fourth International Symposium on Detonation ONR ACR-126, p.14, 1965.
4. Clark, E.L., "Testing of Anti-Armor Devices," BRL-CR-221, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, April 1975. (AD B003825L)
5. Bailey, R.A., Born, D., and Sultanoff, M., "Analysis of the Performance of the Mock-Up Booster Assembly for the Multi-Jet, Shaped Charge, Anti- Tank Mine," BRL-MR-584, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, October 1951, (AD 377333)
6. Cioffi, A.R. and Vincent, A.R., "Preliminary Estimates of the Vulnerability of Light Weight Armored Vehicles to Attack by Antitank Mines," (U), BRL-TN-1197, US Army

Ballistic Research Laboratory, Aberdeen Proving Ground, MD, June 1958 (CONFIDENTIAL). (AD 378697)

7. Johnson, W.E., "Code Correlation Study,: AFWL-TR-70-144, Air Force Weapons Laboratory, Kirtland Air Force Base, NM, April 1971.
8. Miller, J.E., "Preliminary Study of Target Load Prediction by Use of a Hydrodynamic Computer Code," BRL MR-2472, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, April 1975. (AD B0038291)
9. Lambourn, B.D. and Hartley, J.E., "The Calculation of the Hydrodynamic Behavior of Plane One-Dimensional Explosive/Metal System", Fourth International Symposium on Detonation, ONR ACR-126, 1965.
10. Aksu, G. and Ali, R., "Determination of Dynamic Characteristic of Rectangular Plates with Cutouts Using a Finite-Difference Formulation", Journal of Sound and Vibration, Vol.44, pp. 147-158, 1976.
11. Karo, A.M., Walker, F.E., Cunningham, W.G., and Hardy, J.R., "Theoretical Studies of Shock Dynamics in Two-Dimensional Structures", Shock Waves in Condensed Matter - 1981, Nellis, W.J., Seaman, L., and Graham, R.A., eds., AIP Conference Proceedings 78, American Institute of Physics, NY, 1982, p.92.
12. Chanteret, P.Y., "An Analytical Model for Metal Acceleration by Grazing Detonation", Seventh International Symposium on Ballistics, The Hague, Netherlands, p. 515, 1983.
13. Westine, P.S. and Hokanson, J.C., "Procedures to Predict Plate Deformations from Land Mine Explosions" (U), TACOM Technical Report No. 12049, US Army Tank Automotive Command, Warren, MI, August 1975.
14. Haskell, D.F. and Reisinger, M.J., "Armored Vehicle Vulnerability Analysis Model - First Version, Introduction," BRL-R-1857, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, February 1976. (AD 009638L)
15. Haskell, D.F., "Deformation and Fracture of Tank Bottom Hull Plates Subjected to Mine Blast," BRL-R-1587, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, May 1972. (AD 901628)
16. Dehn, J.T., "Models of Explosively Driven Metal, BRL-TR-2626, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, December 1984.
17. Lottero, R.E. and Kimsey, K.D., "A Comparison of Computed Versus Experimental Loading and Response of a Flat Plate Subjected to Mine Blast", Memorandum Report ARBRL-MR-02807, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, January 1978.
18. Gupta, A.D., "Dynamic Analysis of a Flat Plate Subjected to an Explosive Blast", Proceedings of the 1985 ASME International Computers in Engineering Conference and Exhibition, Vol.1, Boston, MA, 4-8 August, 1985.
19. Gupta, A.D., Gregory, F.H., Bitting, R.L. and Bhattacharya, S., "Dynamic Analysis of an Explosively Loaded Hinged Rectangular Plate", Computers and Structures, Vol. 26. No. 1/2 pp 339-344, Pergamon Journals Ltd., UK, August 1987.

20. Engineering Design Handbook, "Explosions in Air, Part One", AMC Pamphlet, AMCP-706-181, Headquarters, US Army Materiel Command, Alexandria, Virginia, July 1974.
21. Soroka, B., "Air Blast Tables for Spherical 50/50 Pentolite Charges at Side-on and Normal Incidence, ARBRL-MR-02975, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, 1979.
22. Goodman, H.J., "Compiled Free-Air Blast Data on Bare Spherical Pentolite, BRL-R-1092, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, 1960.
23. Kingery, C.N. and Bulmash G., " Air Blast Parameter from TNT, Spherical Air Burst and Hemispherical Surface Burst", BRL-TR-2555, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, April 1984.

DEVELOPMENT OF A COMPUTATIONAL METHOD FOR CONVENTIONAL
WEAPONS ANALYSIS OF BURIED STRUCTURES

James T. Baylot
U.S. Army Engineer
Waterways Experiment Station
Vicksburg, MS 39180-6199

ABSTRACT. Current methods, including Army Technical Manual (TM) 5-855-1, of designing buried structures to withstand the effects of the detonation of an earth penetrating conventional weapon, utilize decoupling assumptions in determining the loads on those structures. Free-field stresses and velocities at the structure location are computed and then assumptions are made to convert these stresses to structure loads. A recent series of tests, CONWEB, sponsored by the Defense Nuclear Agency, emphasized the need for better methods of predicting loads on structures.

The most serious drawback of any of the current methods is the decoupling of the free-field stress calculation from the structure loading and structure response calculations. Therefore, an analysis method is needed in which the detonation of the explosive, the propagation of stresses through the soil, the soil-structure-interaction and the structure response are included in a single calculation. The development of this fully coupled analysis method is presented in this paper.

The Finite Element (FE) method is an excellent method for attempting to perform this fully coupled analysis. A constitutive model capable of functioning in the very high stress region near the charge, and a method of stabilizing the solution, without adversely affecting the results, are needed. This model and method were developed and implemented in an FE computer code. A nonreflecting boundary method capable of functioning extremely well in regions of nonlinear response was also developed. This boundary is used to reduce the volume of soil which needs to be modeled and makes the fully coupled analysis method practical.

This paper presents comparisons of the results of analyses performed using this newly developed method, to free-field stress and velocity data measured in tests in sand and clay in the CONWEB test series. Analyses using this method did a good job of predicting those free-field stresses and velocities. Analyses performed to assess the new nonreflecting boundary demonstrated that this boundary functions extremely well. This analysis method has been implemented in an FE computer code capable of modeling the soil-structure-interaction and structure response, and will be extremely useful in developing new methods of designing buried protective structures. The use of this new method will result in tremendous savings in the construction of this type of structure.

BACKGROUND. The military has assets which are critical to its mission and must be capable of surviving a conventional weapons attack. Quite often this objective is met by placing these assets in a buried reinforced concrete structure such as the one shown in Figure 1. These structures are extremely expensive and efforts must be made to minimize costs.

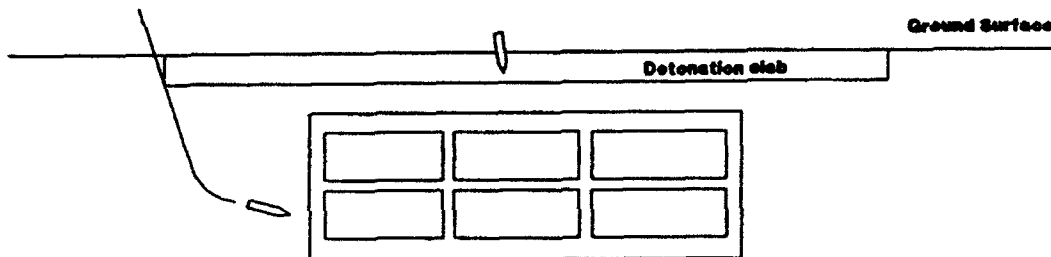


Figure 1. Typical buried hardened structure

Damage to a buried structure from a conventional weapon detonation is usually highly localized. One means of reducing costs is to accept moderate to high levels of localized damage, as long as structural failure does not occur. Accurate design procedures are needed to ensure against failure without being overly conservative; since this increases costs and defeats the purpose of allowing localized damage.

CURRENT ANALYSIS PROCEDURES. Army Technical Manual (TM) 5-855-1 [1] recommends the use of a single-degree-of-freedom (SDOF), spring-mass, system to analyze the structural element being designed. Procedures are outlined to determine the equivalent SDOF system to model one-way and two-way slabs with various end fixity conditions. Equations and figures are presented to determine the capacities of these slabs and for computing stress and velocity histories in the soil. Methods of determining the load on, and response of the structure, as well as recommendations on acceptable response, are provided. The designer can select a trial structural geometry and very quickly evaluate it using the guidance provided in TM5-855-1.

An accurate determination of the structure load is critical to designing the structure. The structure loading is a function of the stresses propagating through the soil and the interaction of the soil with the structure. Thus, the structure loading is coupled to the structure response.

TM5-855-1 recommends a decoupled analysis. The loads on the structure are determined by modifying the free-field stresses (the stresses which would have been present in the soil at the structure location if the structure was absent) and these loads are applied to the SDOF model to determine the structure response. A semi-empirical procedure is used to determine the structure loading from the free-field stresses. Elements of linear elastic wave theory combined with experimental data were used to develop the procedure used to compute the structure load from the free-field stress. The peak structure load is

linearly to the free-field stress in a time equal to twelve transit times of a stress wave through the thickness of the wall or roof slab being analyzed.

An Air Force protective construction manual [2] recommends a procedure very similar to the one used in TM5-855-1. The major difference in the two procedures is the method of computing the loads from the free-field stresses. The decoupling procedure recommended in this manual uses linear-elastic 1-D wave theory to determine the structure loads. Assuming continuity of stress and velocity at the soil-structure interface, the structure interface loading, s_i , is given by:

$$s_i = s_{ff} + qC (V_{ff} - V_s) \quad \text{eqn. 1.}$$

where: s_{ff} is the free-field stress, qC is the acoustic impedance of the soil, and V_{ff} and V_s are the free-field and structure velocities, respectively. This will be referred to as the soil-medium-interaction (SMI) method of determining the load on the structure. This equation is further simplified using the linear elastic 1-D relationship:

$$s_{ff} = qC V_{ff} \quad \text{eqn. 2.}$$

Substituting equation 2 into equation 1 results in the modified SMI (MSMI) equation:

$$s_i = 2 s_{ff} - qC V_s \quad \text{eqn. 3.}$$

Either eqn. 1 or 3 is used to compute the structure loading as long as the result is a positive (compressive) loading on the structure. Since the soil-structure interface cannot support tension, the loading is set equal to zero when eqn. 1 or 3 predicts tension.

The SMI, MSMI, and TM5-855-1 procedures for computing structure load have been evaluated [3] against experiments conducted for the Air

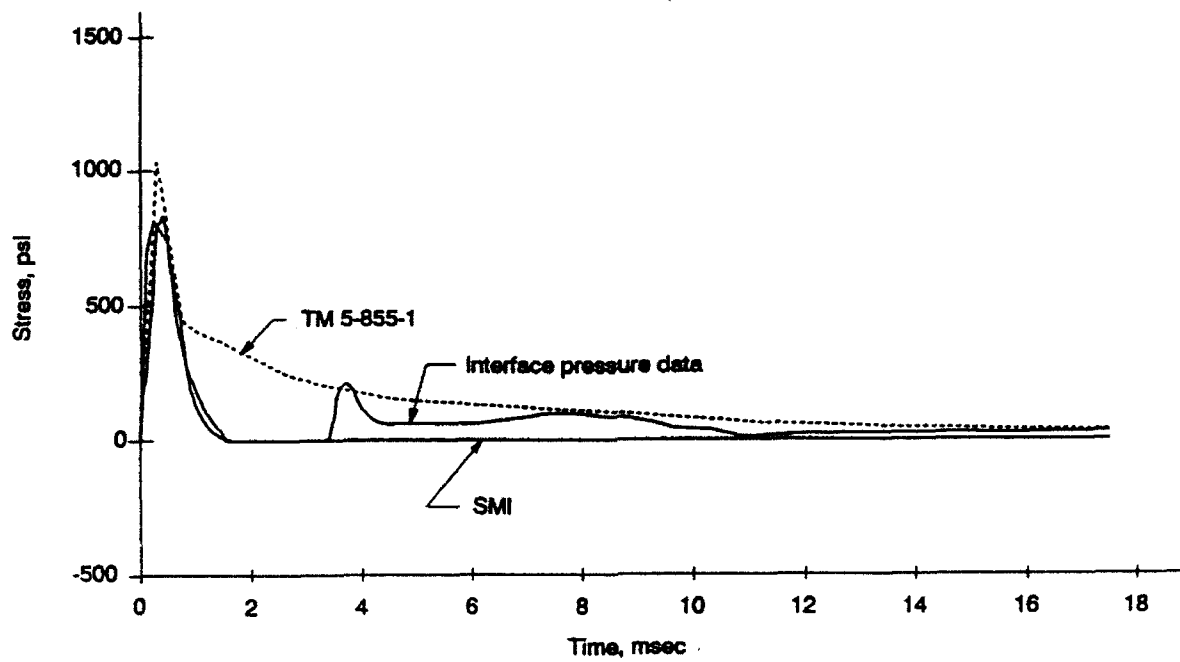


Figure 2. Typical comparison of measured to predicted loads

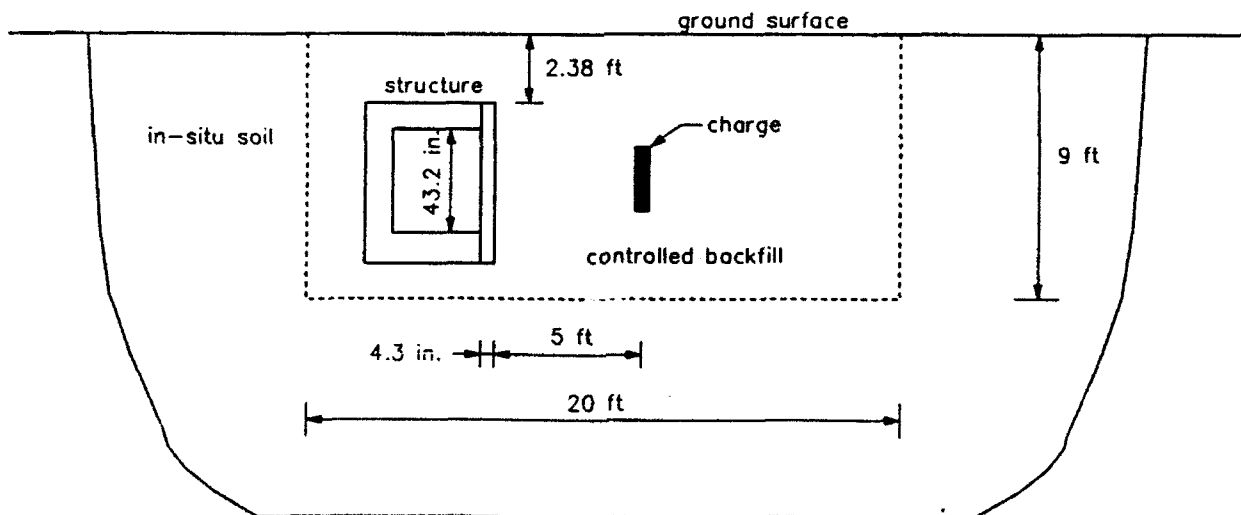


Figure 3. Configuration for CONWEB experiments

Two of these experiments, Figure 3, were identical, except that one was buried in wet clay while the other was buried in dry sand. The structure buried in the wet clay was severely damaged (19 in. deflection on a 43.2 in. span). The damage to the structure in sand was minimal (1.4 in. deflection).

Figure 4 compares the early time loading at the centers of the slabs. This figure shows that the peak loading at the center of the slab buried in clay is much higher than that of the slab buried in sand. The duration of this initial loading is much higher in the sand, resulting in an initial impulse that is approximately the same as for the structure buried in clay. Since the initial loadings are approximately the same, this indicates that the late-time loading on the slab is responsible for the difference in response and late-time loads are important.

These analyses of experiments have indicated that late time loads on buried structures significantly affect the response of the structure and that neither method of predicting loads on the structure adequately predicts these late-time loads. Therefore an improved method of predicting these late-time loads is needed. A study was initiated to determine a better method of accounting for soil-structure-interaction (SSI) effects when predicting structure loads from free-field stresses.

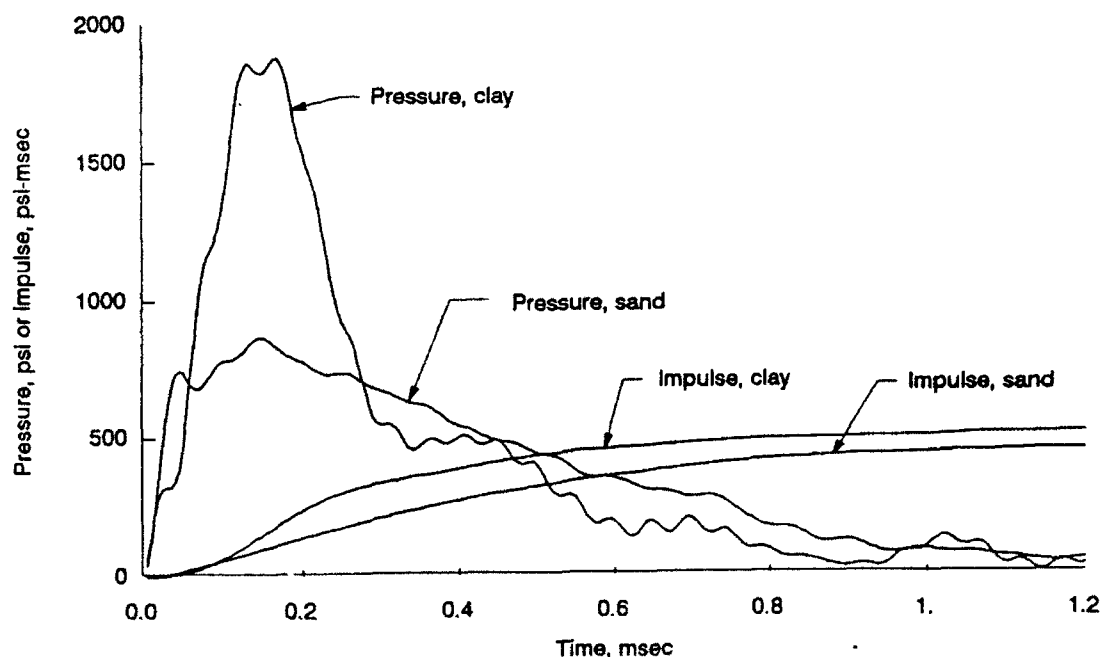


Figure 4. Comparison of early-time loads in clay and sand

NEED FOR FULLY COUPLED ANALYSIS PROCEDURE. In order to develop a better method of decoupling the structure response analysis from the free-field stress analysis, an analysis procedure which does not require decoupling is needed. One possibility is to use a "soil island". The structure and a small "island" of soil around it are modeled in a finite element (FE) analysis. Stresses and/or velocities are input on the boundary of the "soil island".

This method has the advantage that it does not require assumptions to decouple the soil from the structure and it does not require that the detonation of the explosives be included in the analysis. The drawback to this procedure is that it requires the introduction of an artificial boundary, the soil island boundary. This boundary does not affect the very early loads on the structure, but does affect the late-time loading since the boundary must be placed between the explosive and the structure and the explosive is very close to the structure. Since the purpose of this study is to study late-time loads, the soil island approach is not appropriate.

The only way to avoid adding a boundary which affects the loading on the structure is to include the detonation of the explosive and the structure response in the same analysis. Analyses of this type have not been reported in the literature; therefore, a fully-coupled analysis procedure which includes the detonation of the explosive, the propagation of stresses through the soil, the SSI, and the response of the structure was developed.

DEVELOPMENT OF FULLY-COUPLED ANALYSIS PROCEDURE. The development of this fully coupled analysis procedure is presented in this paper. Before this procedure can be used to study SSI, it must be shown that it adequately predicts free-field stresses and velocities. Analyses were performed for comparison with free-field stress and velocity data collected in the CONWEB experiments in wet clay and dry sand. These comparisons are presented in this paper.

The analysis procedure must be capable of adequately modeling the SSI. When a stress wave propagates through the soil to the structure there are several possible occurrences. Initially the soil will load the structure and the structure will start to deform. The structure could deform fast enough so that the soil separates from the structure, and thus the structure is unloaded. Later the soil may catch up with the structure and reload it. It is also possible that the soil may slide on the structure, and the soil may flow past the structure.

There are FE procedures capable of modeling these effects. There are also FE methods for modeling the detonation of explosives, the propagation of the stresses through the soil and the response of the structure. Thus it appears that the FE method is ideal for attempting to develop this fully coupled analysis procedure.

There are several considerations specifically related to the FE method of solving this problem. Constitutive models in FE codes are capable of modeling non-linear behavior such as the behavior of soil. These models are not, in general, capable of functioning in the very

high stress region adjacent to the explosive. Therefore it was expected that the constitutive model would need modification in order to perform the SSI study.

The SSI study will involve the propagation of a very high stress wave through the soil. The extremely short rise time associated with this stress wave cannot be propagated through the FE grid which represents the soil continuum as a series of elements. Attempting to propagate this stress wave through the grid leads to numerical instability. Some method of stabilizing the solution without adversely affecting the response is needed. This is typically accomplished by including artificial viscosity forces. These artificial viscosity forces spread the shock front over several elements and damp out the numerical instabilities. Procedures for stabilizing the solution are built in to FE codes, but these default procedures were not satisfactory for an explosive detonating in soil and a modified method of stabilizing the solution without adversely affecting the solution was developed.

In order to adequately predict the rise times on the free-field stress histories, a fine grid will be needed. The grid size must be limited so that the run times will be reasonable. Nonreflecting boundaries are available to reduce the required grid size. These are boundaries which are used to model an infinite amount of material beyond the boundary.

A nonreflecting boundary was developed by Lysmer and Kuhlemeyer [6] for dynamic loadings. A similar procedure developed by Underwood and Geers [7] is suitable for both static and dynamic loadings. They demonstrated that these boundaries are extremely effective when the boundary is placed in a region of linear elastic response. In another study, [8], they showed that these boundary techniques are not effective when the boundary is placed in a region of highly nonlinear response.

The behavior of the soil will be highly nonlinear in a large region around the charge, and it will be impractical to include all of the soil in the SSI analysis. Therefore, a nonreflecting boundary method capable of functioning in the nonlinear response region of the soil was developed.

The large deformation, large strain, explicit, three-dimensional, FE code, DYNA3D [9] contains the Jones Wilkins Lee (JWL) [10] equation of state which can be used to model the detonation of the explosives. The Cap model [11, 12] is available to model the soil, nonreflecting boundaries are available, and an interface routine which allows separation, recontacting, and sliding with friction of the soil on the structure is available. These features make DYNA3D an excellent candidate for attempting the SSI analyses; therefore, it was selected for this study.

DEVELOPMENT OF HIGH-PRESSURE CONSTITUTIVE MODEL FOR SOIL. Before attempting the SSI analyses it should be demonstrated that the method is capable of adequately predicting free-field stresses and velocities.

This requires that a constitutive model capable of functioning in the very high stress region near the charge and a method of stabilizing the solution be developed and implemented in DYNA3D.

The Cap model is an excellent model for predicting stress wave propagation through soil, but the Cap model in DYNA3D would not converge to a solution for simple test problems using the material properties of the CONWEB clay material. Therefore, another Cap model was obtained [13].

This new Cap model functioned well for both the CONWEB clay and sand materials, but would not function in the very high stress region near the charge. In this new Cap model the volumetric stress strain relationship is divided into the elastic and plastic parts. The change in elastic volumetric strain is the change in pressure divided by the constant bulk modulus, K . At extremely high pressures in soils, the bulk modulus increases significantly with increasing pressures. Ignoring this increase will cause the model to predict unrealistically high volumetric strains. Therefore the Cap model had to be modified so that the bulk modulus varied with pressure.

Static test data [14] at very high pressures have shown that the bulk modulus is adequately represented by the following function:

$$K = K_1 + K_2 P^{K_2} \quad \text{eqn. 4.}$$

Where K_1 , K_2 , and K_2 , are material constants and P is the pressure. The new Cap model was modified to use this function for the bulk modulus. Analyses were then performed to ensure that the modified Cap model combined with the JWL equation of state would satisfactorily predict free-field stresses and velocities.

DEVELOPMENT OF STABILIZATION METHOD. In the initial attempts to validate the Cap model, the FE analyses were unstable. DYNA3D [15] uses linear and quadratic artificial viscosity forces to spread the shock front over several elements. An artificial force, f , is introduced for elements which are compressing. This is the method developed by von Neumann and Richtmyer [16] and it has been demonstrated that this method stabilizes the solution, but does not perturb the solution away from the shock front. The artificial force is given by:

$$f = q_1 \left| \dot{\epsilon}_{kk} \right| (Q_1 l \left| \dot{\epsilon}_{kk} \right| + Q_2 c) \quad \text{eqn. 5.}$$

Where Q_1 , and Q_2 are dimensionless constants, q is the mass density, l is the characteristic length of the element, c is the sound speed for the material, and $\left| \dot{\epsilon}_{kk} \right|$ is the magnitude of the volumetric strain rate. For relatively uniform grids, it is appropriate to use the cube root of the element volume for l , and this is done in DYNA3D. Q_1 and Q_2 default to 1.5 and 0.06, respectively, in DYNA3D.

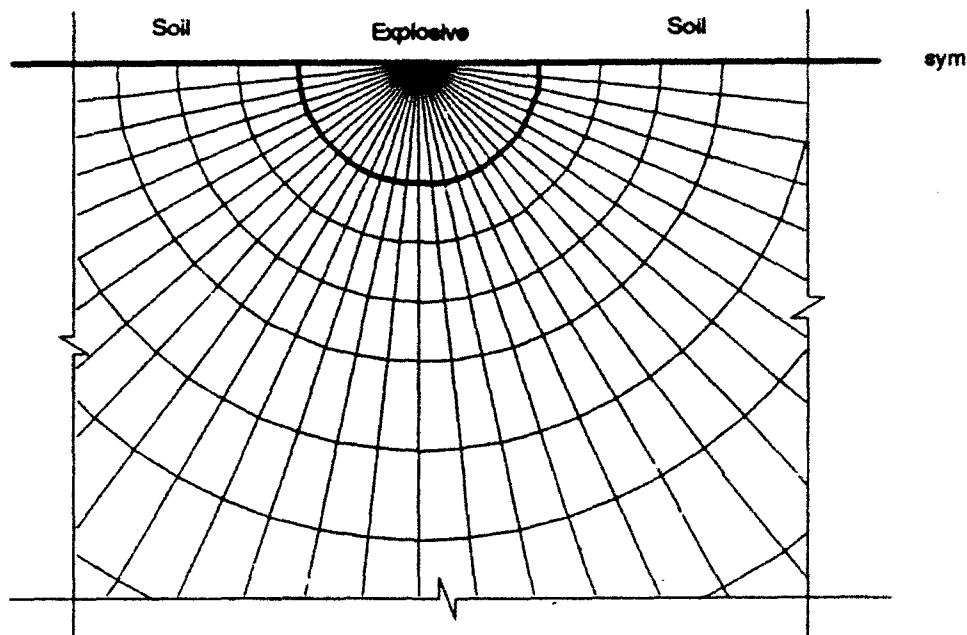


Figure 5. Grid for explosives and surrounding soil

In the SSI analyses a 2-D cylindrical grid, as shown in Figure 5, will be used to model the charge and surrounding soil. Because of the shape of the elements and the reduction of stresses as the shock propagates away from the charge, it is more appropriate to use viscosity forces which are proportional to the radial dimension of the element, rather than the cube root of the volume.

Using the cube root of the volume will result in artificial viscosity forces which are too low near the charge. This effect can be offset by using values of the constants Q_1 and Q_2 which are higher than the default values. At distances away further away from the charge the artificial viscosity forces will be too high, and will affect the predicted stress and velocity histories. This effect can be reduced by modifying the method of computing the characteristic length of the element, or by using artificial viscosity terms which vary with distance from the charge. The latter method was easier to implement, therefore it was used.

COMPARISONS TO EXPERIMENTAL FREE-FIELD DATA. One-dimensional spherical analyses were performed for comparison with the free-field data in the CONWEB experiments in clay and sand. A complete description of these analyses has been reported [17]. In these experiments free-field stresses and accelerations were measured at 3, 4, 5, 6, and 7 ft from the explosive. The acceleration histories were integrated to obtain the free-field velocities.

Initially the analyses in clay were attempted using the default values of the artificial viscosity coefficients, but these analyses were unstable. The lowest viscosity coefficients which produced a stable solution were 5 times the default values. The predicted stresses for this analysis are compared to data in Figure 6. This figure shows that stresses are predicted extremely well at the 3 and 4 ft ranges. At the other ranges the peak stresses are well predicted, but the rise times are too long and the late-time stresses are too high.

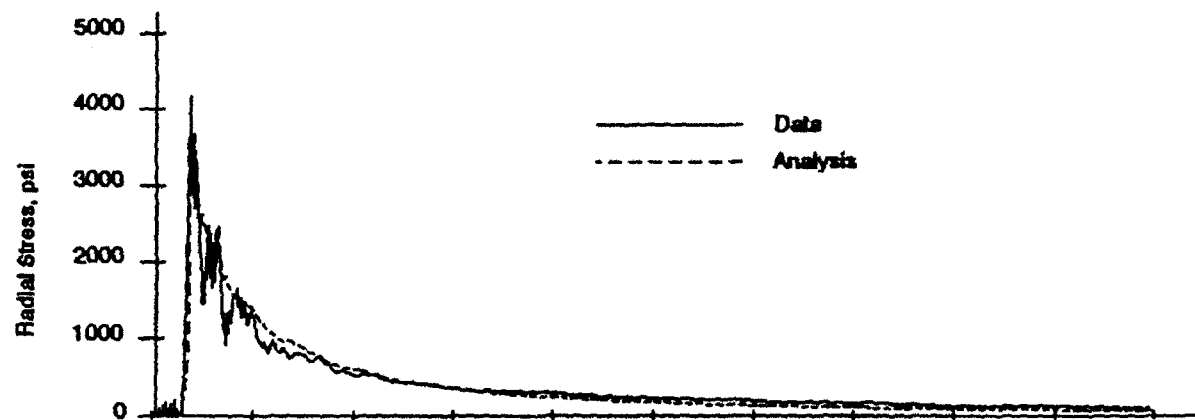
Figure 7 compares the free-field velocities to the data for this experiment. The velocity histories also agree reasonably well with the data. The late-time velocities are slightly overpredicted and the peak velocities at the farther ranges are overpredicted. Rise times to maximum velocities are overpredicted. The overpredicted stresses at the farther ranges are due largely to the free surface which is present in the experiment, but not included in the analysis.

The large rise times of the stress and velocity histories, as compared to the data, are due to the method which DYNA3D uses to stabilize the solution. Rise times which are too long could affect the loads transmitted to the structure; therefore it is important to model them reasonably well.

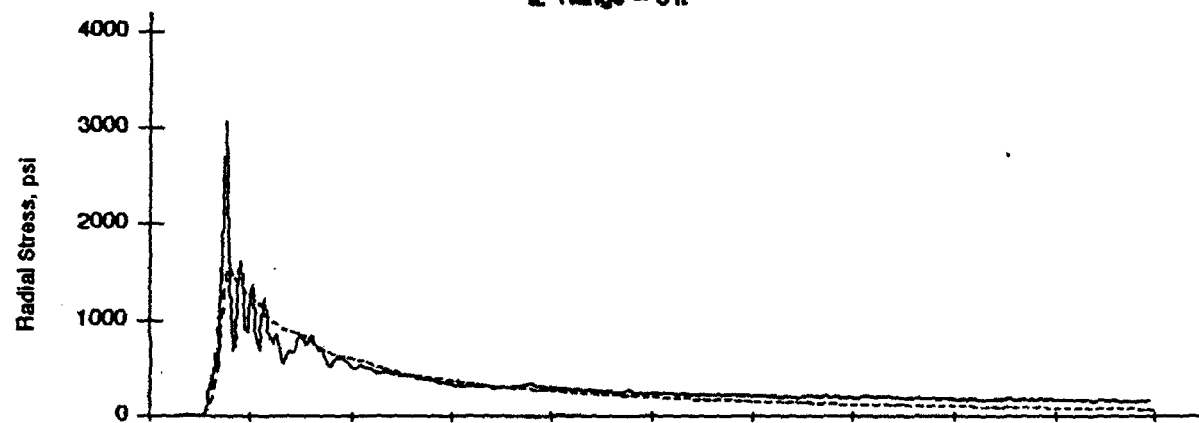
This analysis was repeated using the newly developed method of stabilizing the solution. This did not affect the stress and velocity histories at 3 and 4 ft but did affect those at the farther ranges. The stress and velocity histories at the 5, 6, and 7 ft ranges are compared to the original analysis in Figures 8 and 9, respectively. These figures show that the stress histories do not change significantly, except for the rise times which are significantly improved by using the new stabilization method. The latter analysis does a much better job of matching the rise times of the experimental data.

Analyses were also performed for comparison with the CONWEB sand experiment. These analyses could not be performed using the default viscosities. Viscosity coefficients of 3 times the default values were needed to stabilize the solution. Free-field stress and velocity histories were predicted reasonably well. In general, rise times were too long and arrival was too late. Late time stresses and were overpredicted and late-time velocities were underpredicted. The late-time overpredicted stresses and underpredicted velocities are due largely to the effects of the free surface and the finite test pit. These effects were not included in the analysis.

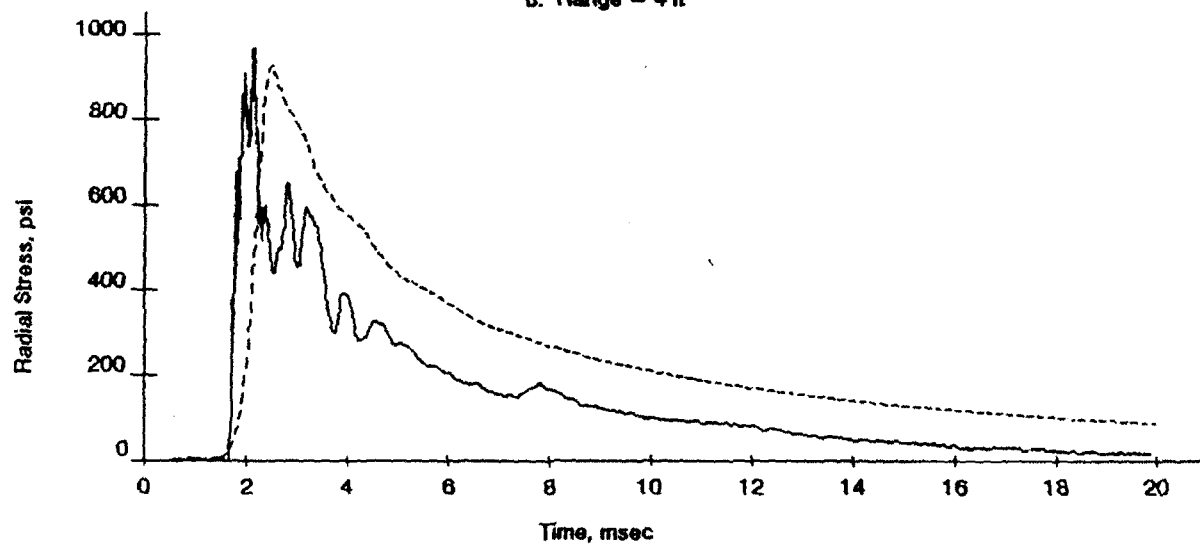
Arrival times and rise times at the 6 and 7 ft ranges were improved significantly by using the modified method of stabilizing the solution. This does increase the predicted peak pressures and velocities but the overall effect of using the newly developed stabilization procedure is to improve agreement between the analysis and the experimental data.



a. Range = 3 ft



b. Range = 4 ft



c. Range = 5 ft

Figure 6. Stresses in clay (continued)

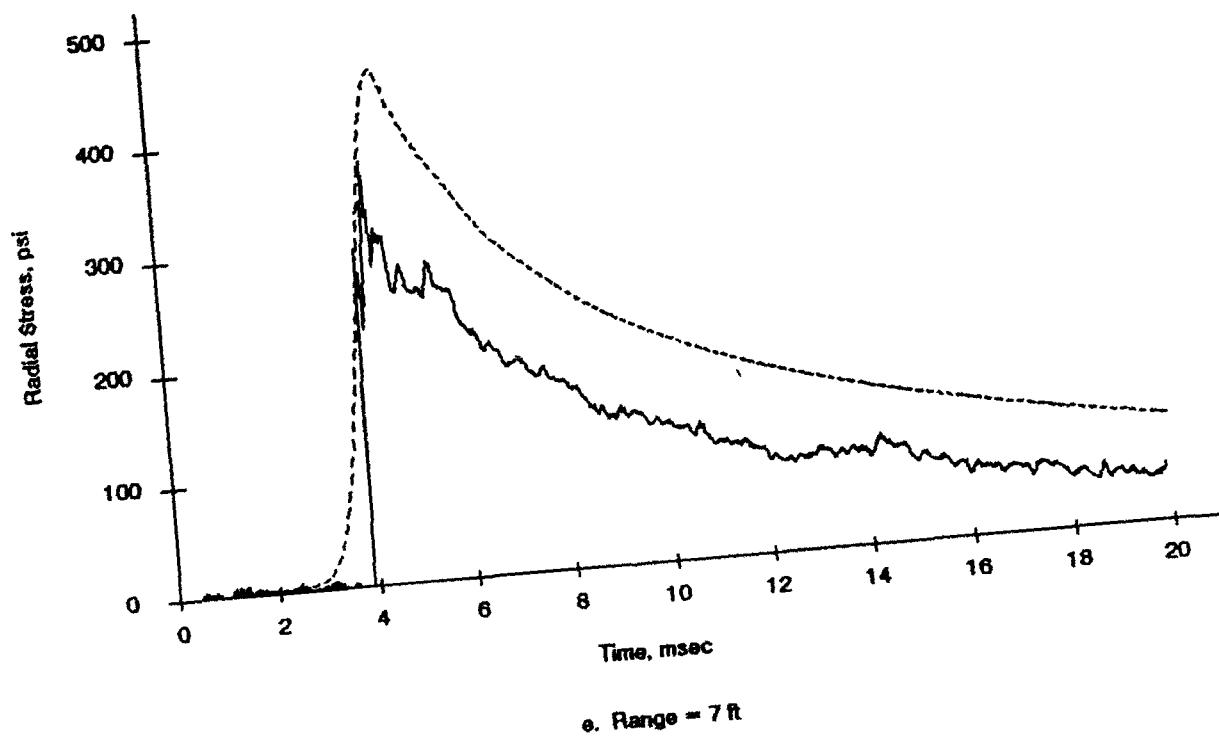
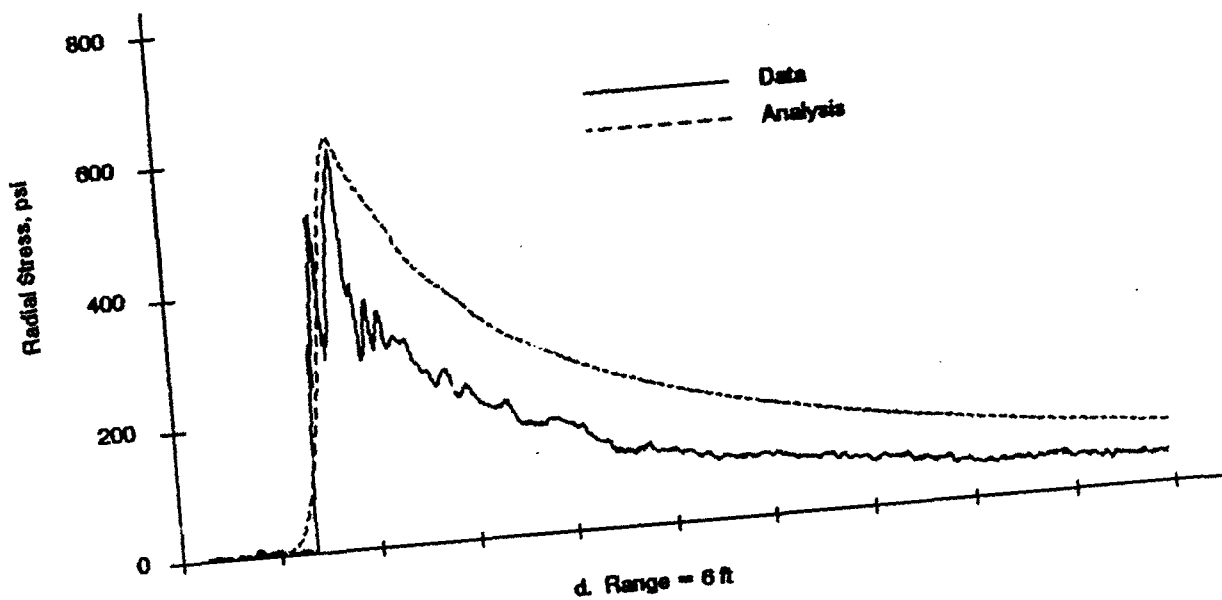


Figure 6. (concluded)

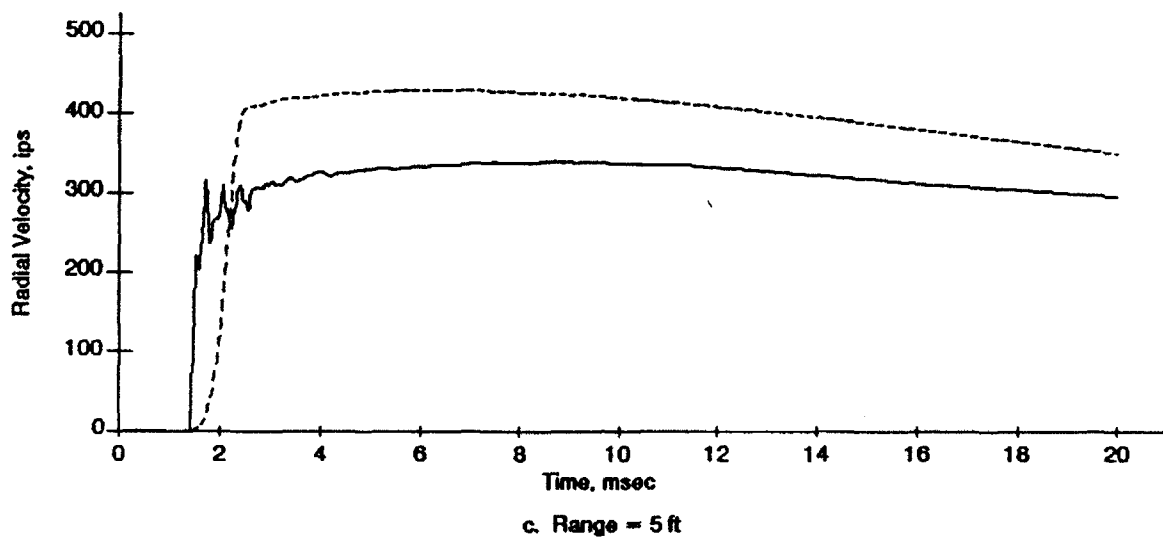
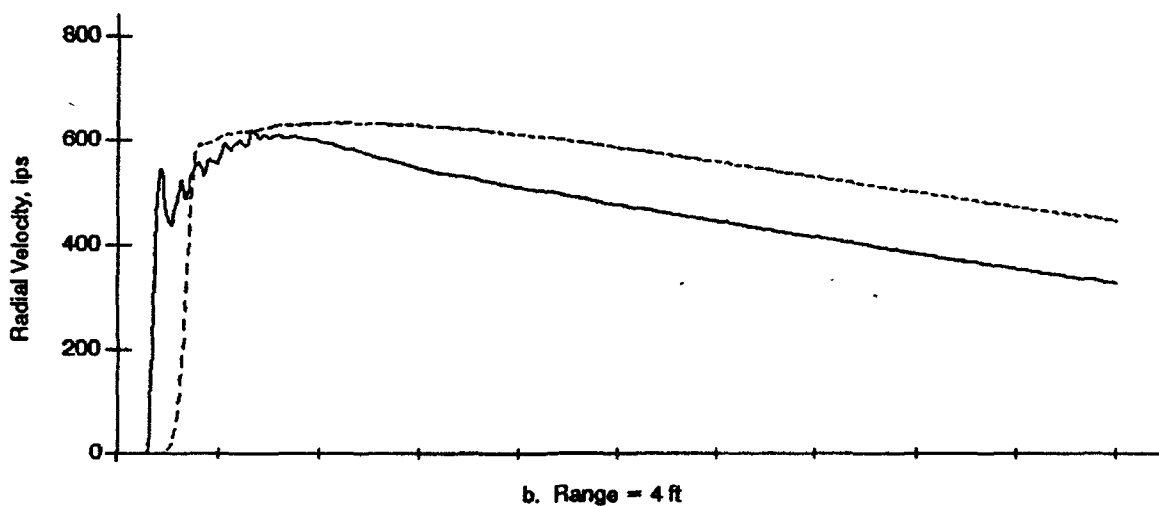
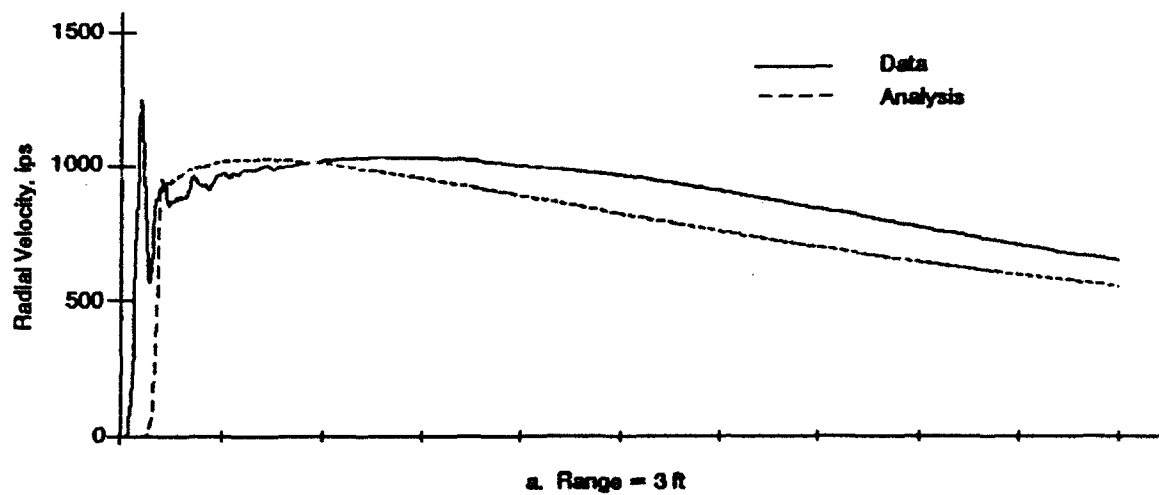


Figure 7. Velocities in clay (continued)

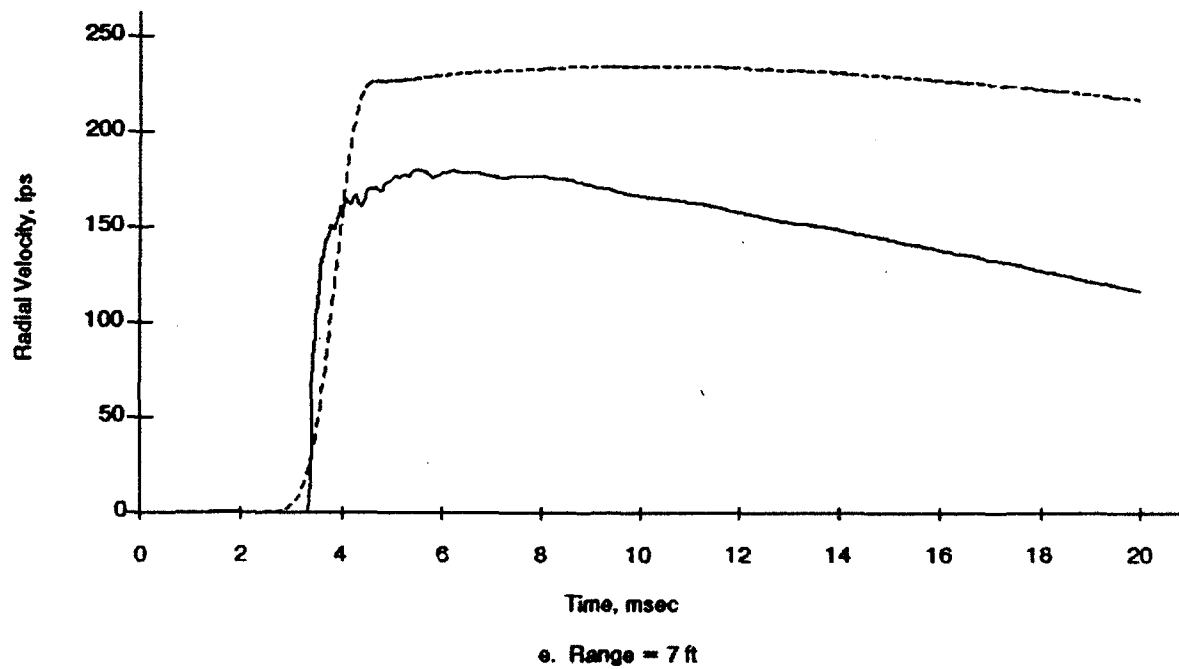
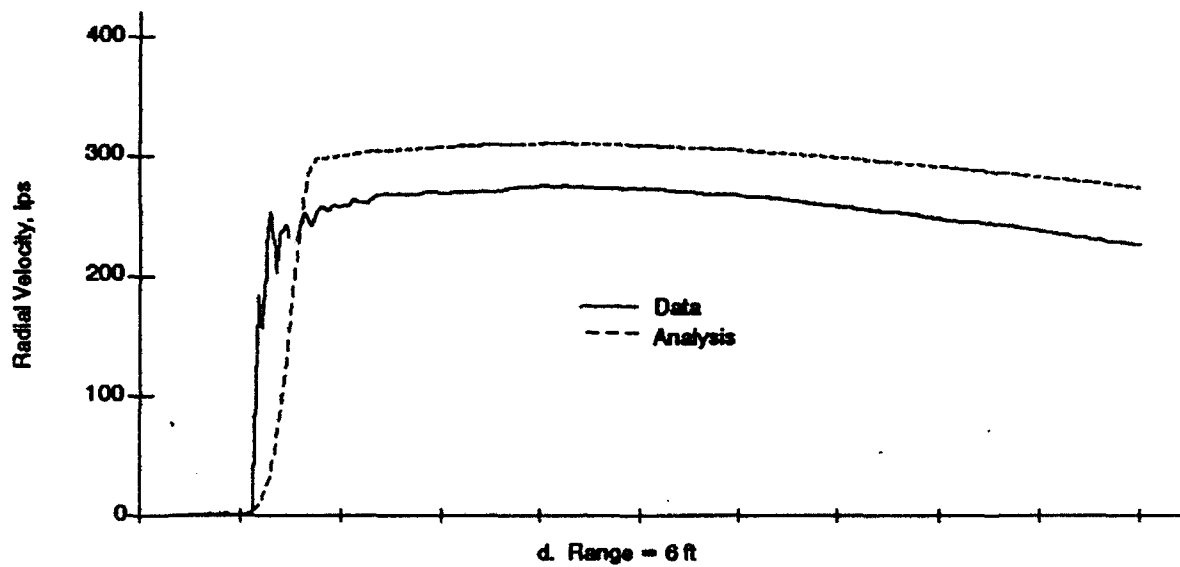


Figure 7. (concluded)

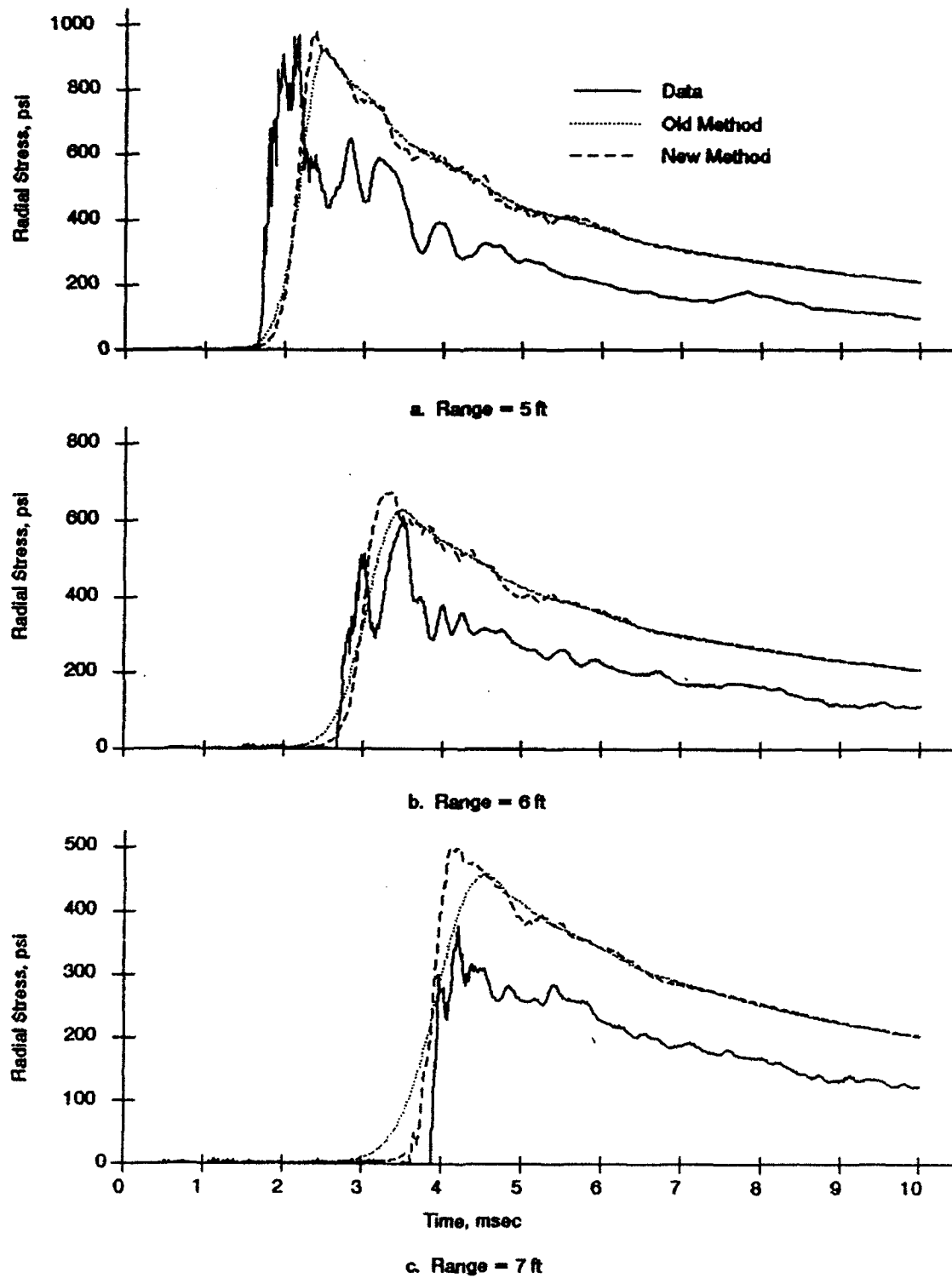


Figure 8. Stresses in clay using new stabilization method

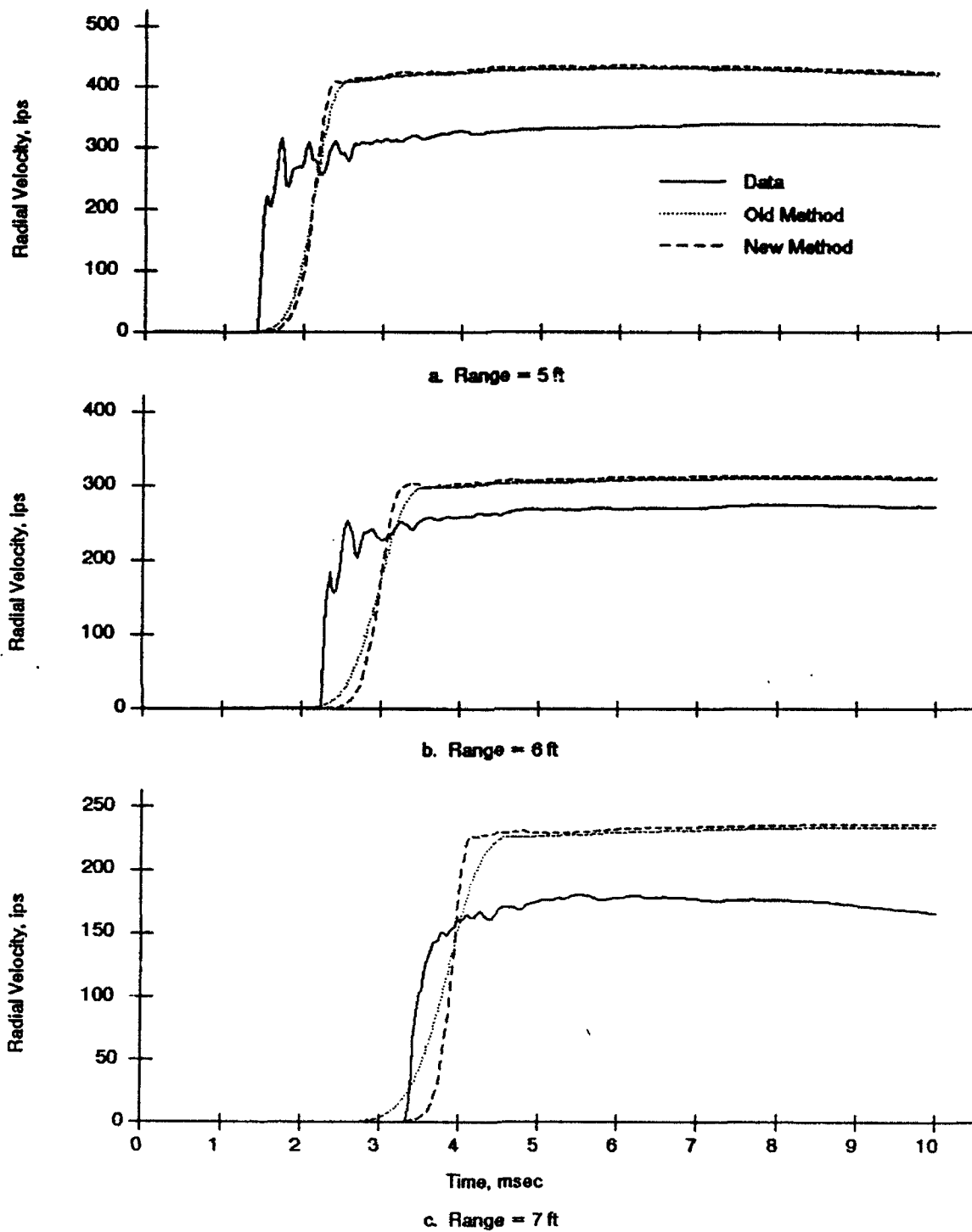


Figure 9. Velocities in clay using new stabilization method

NONREFLECTING BOUNDARIES. Figure 10 shows the results of several analyses performed using a 1 in. radius charge in clay in a 1-D cylindrical geometry. In one case the grid boundary was placed at 40 ft from the charge, and the effects of this boundary do not appear in the time period shown. In the other histories, three different types of boundaries were placed at 10 ft from the center of the charge. This figure shows that using the nonreflecting boundary installed in DYNA3D produces about the same stress history as providing a fixed boundary at the same location.

This boundary is based on the elastic properties of the material and is obviously not effective when placed in a nonlinear response region. The nonreflecting boundary in DYNA3D is based on the work of Lysmer and Kuhlemeyer [7], and amounts to placing viscous springs at the boundary to simulate the material which would have been on the other side of the boundary. The constants for these viscous springs are based on the elastic properties of the material.

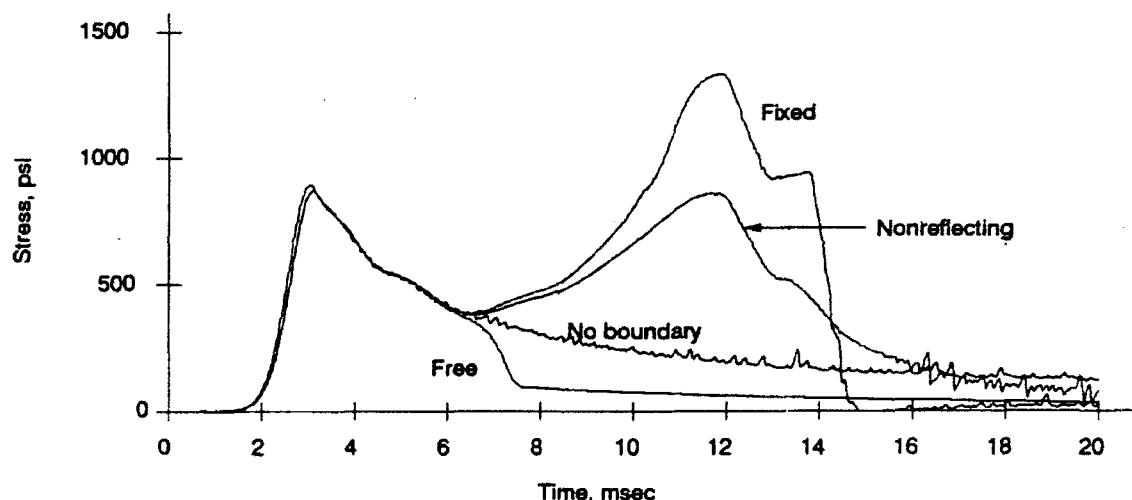


Figure 10. Effect of boundaries on stress history

Figure 11 shows the volumetric stress-strain relationship for the CONWEB clay. The solid curve in this figure includes both elastic and inelastic response. This curve shows a significant stiffening of the material at strains above 4 percent. At this time, all of the air voids have been compressed out of the soil and the soil behaves as an elastic material. The material will unload parallel to the elastic curve. The dotted curve shows the elastic pressure volumetric strain relationship at low pressures. At strains above 4 percent the total strain curve is parallel to the elastic strain curve.

The nonreflecting boundary is based on this elastic curve, and the boundary is too stiff until the strains reach 4 percent. For a monotonically decreasing stress pulse similar to the ones shown in Figure 6, propagating into an infinite amount of soil, there will always be material in front of the shock which is strained to less than 4 percent, therefore the nonreflecting boundary will always be too stiff.

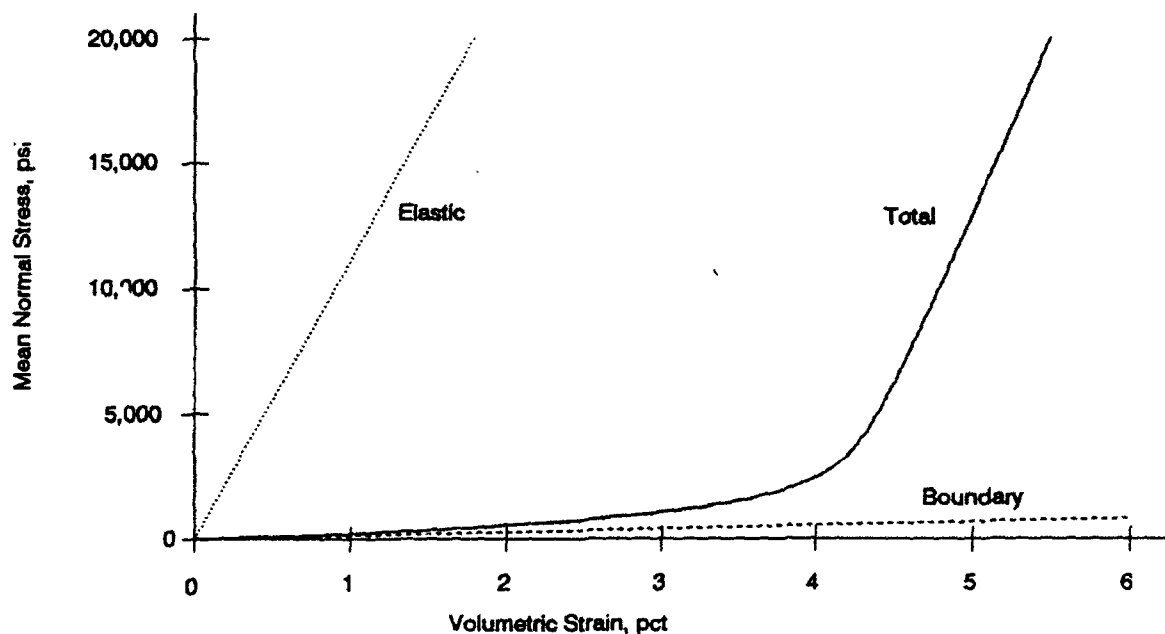


Figure 11. Hydrostatic response of clay

The nonreflecting boundary was modified to reflect the total stress-strain behavior of the material instead of the elastic behavior. This was done by using a stress-strain relationship like the one shown as the dashed curve in Figure 11. This was accomplished by modifying the subroutine which provides the material property information to the nonreflecting boundary subroutine. Using this method, the nonreflecting boundary problem can be corrected without affecting the material constants used for the wave propagation analysis.

Because this problem involves a single outgoing monotonically decaying stress pulse, this nonreflecting boundary should perform satisfactorily. A comparison of analysis results with the boundary far away to those using the modified nonreflecting boundary is presented in Figures 12 and 13 for stresses and velocities, respectively. These figures show that the modified nonreflecting boundary is extremely effective.

Similar results were obtained when the modified nonreflecting boundary was used in the sand backfill material.

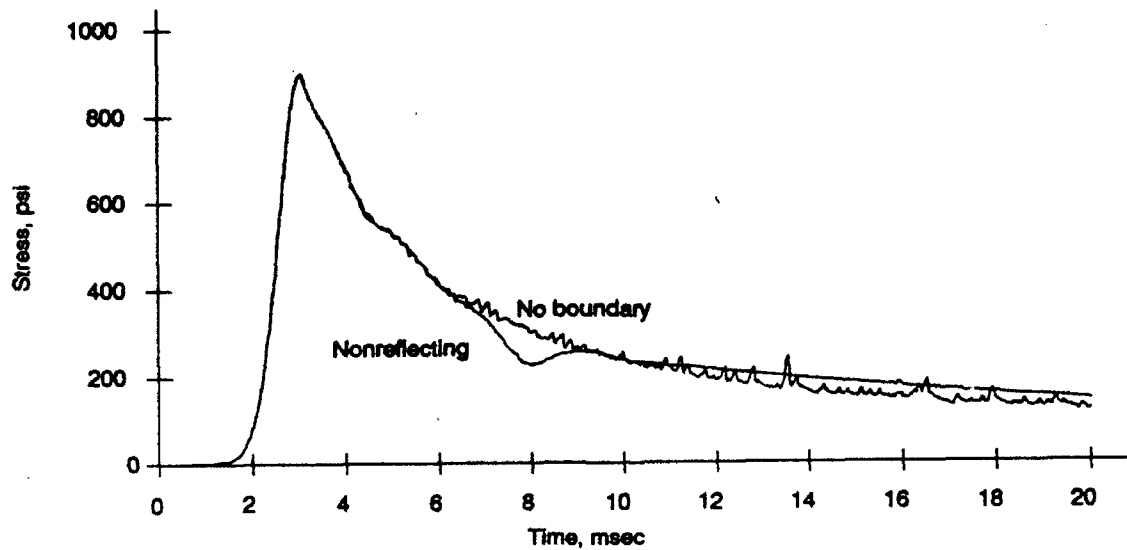


Figure 12. Stress histories using new nonreflecting boundary

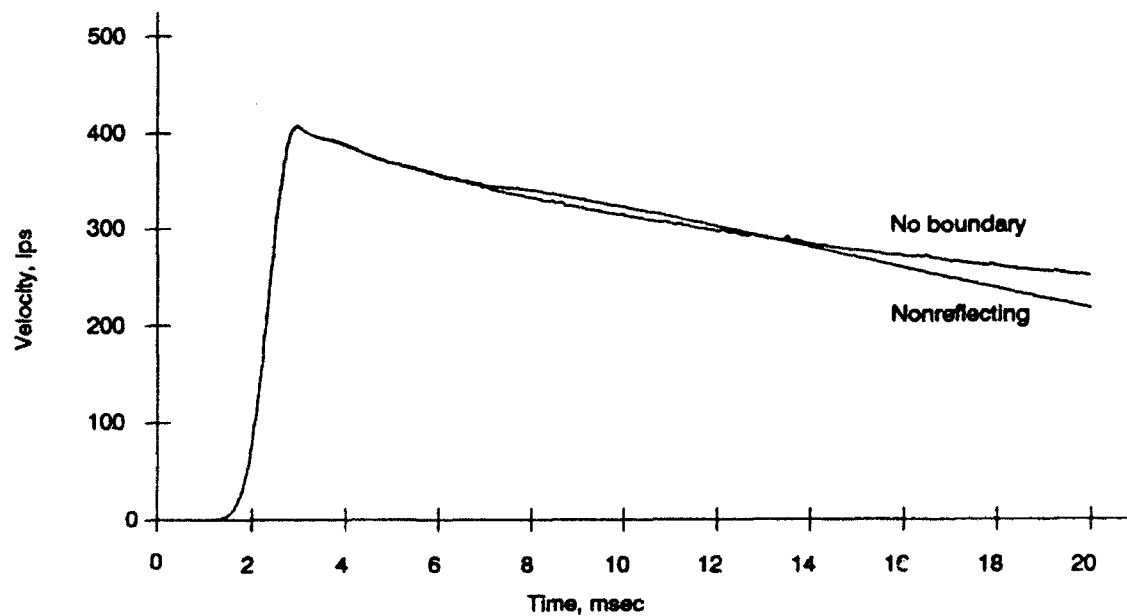


Figure 13. Velocity histories using new nonreflecting boundary

SUMMARY. Accurate design procedures are needed so that the costs of buried hardened structures can be reduced. The most serious drawback of current analysis methods is the method used to decouple structural response calculations from free-field stress calculations. A fully coupled analysis procedure which includes the detonation of the explosive, the propagation of stresses through soil, the soil-structure interaction, and the response of the structure was needed.

It was decided that the FE method was appropriate for attempting these SSI analyses. There were several problems with existing FE methodology which had to be addressed before the FE method could be used for the SSI study. Typical constitutive models for soil do not function well in the extremely high stress region adjacent to the charge. A modified Cap model capable of performing immediately adjacent to the charge was developed. The same constitutive model can also be used at much lower stress levels far away from the charge.

The methods typically used in FE codes to stabilize the solution for shock wave propagation were not adequate for computing stress wave propagation in cylindrical and spherical grid geometries near the model of a explosive detonation. A method which stabilized the solution, without adversely affecting the problem solution was developed.

Analyses were performed to assess the modified Cap model and stabilization method. The analysis results using these modifications agreed extremely well with test data.

A nonreflecting boundary is required to reduce the grid size so that run times will be reasonable. Currently available boundaries function well in regions of elastic response but do not function well in regions of significant nonlinear inelastic response. In the SSI study, the boundaries must be placed in regions of significant nonlinear response. Therefore, a nonreflecting boundary routine which functions well in regions of nonlinear behavior was developed. Analyses performed using this modified boundary showed that the stress and velocity histories were very similar to analyses in which the boundary was far away.

CONCLUSIONS. The modified Cap model along with the JWL equation of state and the new method of stabilizing the solution functioned extremely well. Predicted stresses and velocities agreed reasonably well with experimental data in both wet clay and dry sand. The newly developed nonreflecting boundary worked extremely well when placed in a region of highly nonlinear response in the soil. These methods were verified in wet clay and in dry sand. The response of most soils will fall between these two. Therefore, these methods should perform equally as well for these intermediate soils.

With the new methods and procedures presented in this paper, a fully coupled analysis, including the detonation of the charge, the propagation of stresses through the soil, the soil-structure interaction and the response of the structure is possible. This fully coupled analysis procedure can be used to develop better methods of predicting loads on buried structures. The use of more accurate procedures for

predicting loads on structures will result in tremendous cost savings in the construction of hardened structures.

This method can be used to analyze structures when other methods, such as the soil island approach are not appropriate. The soil island approach can only be used if the bomb is relatively far away from the structure, or if late-time loads are unimportant. The newly developed method can be used to model both early- and late-time behavior. The new method is appropriate for charge standoffs of only several charge diameters. Therefore this method is applicable for almost all conventional weapons analyses except for contact bursts. Since this is an FE method, a wide variety of structure geometries can be modeled.

ACKNOWLEDGMENTS. This research was sponsored by and conducted at the U.S. Army Engineer Waterways Experiment Station. I gratefully acknowledge permission from the Chief of Engineers to publish this paper.

REFERENCES.

1. Department of the Army Technical Manual 5-855-1, Fundamentals of Protective Design for Conventional Weapons, November, 1986.
2. Drake, J. L., and others, Protective Construction Design Manual: Loads on Structures (Section VIII), ESL-TR-87-57, Air Force Engineering and Services Center, Engineering and Services Laboratory, Tyndall Air Force Base, FL., November, 1989.
3. Baylot, J. T., and Hayes, P. G., "Ground Shock Loads on Buried Structures," Proceedings of the 60th Shock and Vibration Symposium, David Taylor Research Center, Bethesda, MD, Vol I, November, 1989, pp. 331-347.
4. Baylot, J. T., Kiger, S. A., Marchand, K. A., and Painter, J. T., "Response of Buried Structures to Earth Penetrating Conventional Weapons," ESL-TR-85-09, Air Force Engineering and Services Laboratory, Air Force Engineering and Services Center, Tyndall Air Base, FL, November, 1985.
5. Hayes, P. G., "Backfill Effects on Response of Buried Reinforced Concrete Slabs," Technical Report SL-89-18, U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS, September, 1989.
6. Lysmer, J., and Kuhlemeyer, R. L., "Finite Dynamic Model for Infinite Media," ASCE Journal of the Engineering Mechanics Division, Vol 95, No. EM4, August, 1969, pp. 859-877.
7. Underwood, P. G., and Geers, T. L., "Doubly Asymptotic Boundary-Element Analysis of Dynamic Soil-Structure Interaction," DNA Report 4512T, Defense Nuclear Agency, Washington, DC, March, 1978.
8. Underwood, P. G., and Geers, T. L., "Doubly Asymptotic Boundary-Element Analysis of Non-Linear Soil-Structure Interaction," DNA Report 4953F, Defense Nuclear Agency, Washington, DC, June, 1979.

9. Hallquist, J. O., and Benson, D. J., "DYNA3D user's Manual (Nonlinear Dynamic Analysis of Structures in Three Dimensions)," Report UCID-19592, Lawrence Livermore National Laboratory, University of California, Berkeley, CA, July, 1987.
10. Dobratz, B. M., "LLNL Explosives Handbook, Properties of Chemical Explosives and Explosive Simulants," Report UCRL-52997, Lawrence Livermore National Laboratory, University of California, Berkeley, CA, 1981.
11. Sandler, I. S., and Rubin, D., "An Algorithm and a Modular Subroutine for the Cap Model," International Journal of Numerical Analysis Methods in Geomechanics, 3, 1979, pp 173-186.
12. Simo, J. C., Ju, J. W., Pister, K. S., and Taylor, R. L., "An Assessment of the Cap Model: Consistent Return Algorithms and Rate Dependent Extension," Report No. UCB/SES-85/5, Department of Civil Engineering, University of California, Berkeley, CA, May, 1985.
13. Pelessone, D., "A Modified Formulation of the Cap Model," Draft Report GA-C19579, General Atomics, San Diego, CA, prepared for the Defense Nuclear Agency, Washington, DC, January, 1989.
14. Personal Communication with Dr. Jon Windham, research Civil Engineer, U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS, November 1, 1989.
15. Hallquist, J. O., "Theoretical Manual for DYNA3D," Report UCID-19401, Lawrence Livermore National Laboratory, University of California, Berkeley, CA, March, 1983.
16. von Neumann, J., and Richtmyer, R. D., "A Method for the Numerical Calculation of Hydrodynamical Shocks," Journal Applied Physics, No. 21, 1950, p 232.
17. Baylot, J. T., "Parameters Affecting Loads on Buried Structures Subjected to Localized Blast Effects," Technical Report SL-92-9, U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS, April, 1992.

ADAPTIVE GRIDS FOR THE HULL HYDRODYNAMICS CODE

C. Wayne Mastin
U. S. Army Corps of Engineers Waterways Experiment Station
and
Mississippi State University

Abstract

The numerical modeling of an explosion and the effects of its pressure loading on surrounding structures involve several different length scales. The failure of a grid to resolve these length scales can lead to an inaccurate solution which is characterized by grid orientation effects and nonphysical oscillations near wavefronts. An adaptive grid which concentrates grid points in regions where solution gradients are large can minimize these types of errors. As an application of the use of adaptive grids, an automatic rezone scheme has been developed for the two-dimensional Cartesian and axisymmetric Eulerian finite-difference algorithms in the HULL hydrodynamics code. The rezone algorithm is completely automated with the grid point locations based on values of the pressure and its derivatives. Sample computations demonstrate how adaptive rezoning can be used to calculate more realistic solutions without increasing the total number of grid points. The examples include the simulation of atmospheric, underwater, and underground explosions.

1 INTRODUCTION

The success or failure of a numerical simulation often depends on the grid that is used. The failure of a grid to resolve all of the significant length scales in a problem will lead to both an inaccurate and an unrealistic solution. This is especially true in the modeling of explosions where large solution gradients must be resolved by the numerical algorithm. An insufficient number of grid points in the region near a shock wave leads to either oscillations or smearing in the solution. A grid with uniform spacing in each coordinate direction could be used for solving problems of this type, but that would be wasteful since there are typically large regions where the solution is nearly constant. It would be desirable to have a grid with a high concentration of grid points in regions where the solution gradients are large and very few points where the solution is nearly constant. However, this cannot be done *a priori* since the shock wave location is determined by the solution. What is needed is an adaptive rezoning capability that regenerates the grid during the computational procedure based on the current values of the numerical solution.

The purpose of this report is to apply one-dimensional concepts of equidistribution to rezone Cartesian grids for solving multi-dimensional problems. The equidistribution principle has been used extensively for solving one-dimensional problems. One of the earlier surveys appeared in the paper by Russell and Christiansen [1]. With a multi-dimensional Cartesian grid system, the rezoning can be done directly without iteration, which is often not practical when applying analogous methods to general curvilinear coordinate systems. The adaptive rezoning method has been used in the solution of several problems involving

the numerical simulation of explosions. The method has worked well on these problems, which were solved using the HULL hydrodynamics code, and could be applied to other problems where the grid must be composed of horizontal and vertical lines. The HULL code has several existing rezone options, but none of them was applicable to the solution of explosion problems.

2 REZONE ALGORITHM

Let us begin by considering a two-dimensional problem which is to be solved using a cell-centered finite volume scheme. The finite difference grid is composed of vertical and horizontal lines

$$\begin{aligned}x &= x_i, i = 0, 1, \dots, i_{\max} \\y &= y_j, j = 0, 1, \dots, j_{\max}.\end{aligned}$$

The spatial intervals are

$$\begin{aligned}\Delta x_i &= x_i - x_{i-1}, i = 1, \dots, i_{\max} \\ \Delta y_j &= y_j - y_{j-1}, j = 1, \dots, j_{\max}.\end{aligned}$$

The values of the numerical solution are defined at the cell centers so that for a typical solution variable p , the value in cell i, j is

$$p_{i,j} = p\left(\frac{1}{2}(x_i + x_{i-1}), \frac{1}{2}(y_j + y_{j-1})\right).$$

Now a new grid is to be constructed based on these solution values defined on an existing grid. An adaptive grid must sense variations in the solution and adjust the grid accordingly. Since there may be several solution variables, more than one adaptive grid could be constructed. However, in explosives problems, the pressure is a key variable in assessing blast effects. Thus, the rezoning algorithm used in the later examples is based on the variations in the pressure.

Since the same scheme is used to redistribute points in each coordinate direction, only the redistribution along the x -axis will be explained in detail. Now any distribution of grid points can be viewed as being defined by a continuous mapping from the parametric interval $[0, i_{\max}]$ onto the coordinate interval $[x_0, x_{i_{\max}}]$. The parametric variable will be denoted by ξ and the grid points x_i are the images of the points $\xi = i$ under the mapping from parametric to coordinate variables. Under this terminology, a uniform grid would be produced whenever the mapping is a linear function. But linear functions are the general solution of the second order differential equation

$$\frac{d^2 x}{d\xi^2} = 0.$$

The particular solution which gives rise to the uniform grid would also satisfy the boundary conditions

$$\begin{aligned}x(0) &= x_0 \\ x(i_{\max}) &= x_{i_{\max}}.\end{aligned}$$

By considering more general boundary value problems, methods can be developed for generating nonuniform grids. To be more specific, nonuniform grids will be constructed as solutions of the second order equation

$$\frac{d}{d\xi}\left(w \frac{dx}{d\xi}\right) = 0$$

where $w = w(\xi)$ is a weight function used to control the distribution of grid points. This equation can be integrated once to give the first order equation

$$w \frac{dx}{d\xi} = \text{constant}.$$

The following rezone scheme will be based on a discretization of this first order equation.

Having motivated the concept of grid distribution from a continuous model, the discrete problem of generating the grid point locations will now be addressed. The first consideration will be the selection of a weight function to be used to control the grid spacing. Each weight function will control spacing in only one direction and must be a function of a single variable. Therefore we will start out by defining a function P by the formula

$$P_i = \max \{p_{i,j} : 1 \leq j \leq j_{\max}\}. \quad (1)$$

In order that the grid spacing be influenced by the values of both p and its derivatives, the weight function will be given as

$$w_i = c_0|P_i| + c_1|P'_i| + c_2|P''_i|, \quad (2)$$

where P' and P'' denote the finite difference approximations of first and second derivatives. The coefficients c_i are included to add flexibility to the scheme. If c_0 is the dominant coefficient, then the grid spacing will be smallest where the function P is greatest. The c_1 term causes grid points to cluster where the derivative is largest, such as near shock waves. The c_2 term would cluster points where large changes in the derivative occur, such as near oscillations in the numerical solution. Nearly all numerical algorithms work best when there is a smooth change in grid spacing. Since w is defined in terms of P and its differences, it may change drastically from point to point and from time step to time step. In such cases it is advisable to smooth the function P before computing w . A diffusive smoother is used which is based on the numerical solution of the heat equation on a nonuniform grid.

The motivation for the smoothing formula will be accomplished by considering again functions of a continuous variable. Suppose that $P(x)$ is the initial value for the solution of the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}$$

This equation can be written in terms of the parametric variable ξ as

$$\left(\frac{\partial x}{\partial \xi}\right)^3 \frac{\partial u}{\partial t} = \frac{\partial x}{\partial \xi} \frac{\partial^2 u}{\partial \xi^2} - \frac{\partial^2 x}{\partial \xi^2} \frac{\partial u}{\partial \xi} \quad (3)$$

Now consider the discretization of the above parabolic equation with the time step chosen locally so that the maximum stable time step is used at each point. This is the procedure

that was used to arrive at the smoothing formula used in our algorithm. The value of P on the right hand side of the following equation is the input grid function and the value on the left is the smoothed value computed by taking one time step in the numerical solution of equation (3).

$$P_i = P_i + \tau(P_{i+1} + P_{i-1} - 2P_i - \frac{1}{2}\sigma(P_{i+1} - P_{i-1}))$$

where

$$\tau = \frac{1}{4} \min \left\{ \frac{1}{|\sigma|}, 1 \right\}$$

$$\sigma = 2 \frac{\Delta x_{i+1} - \Delta x_{i-1}}{\Delta x_{i+1} + \Delta x_{i-1} + 2\Delta x_i}$$

The parameter τ is chosen from the stability analysis. Note that σ is due to the nonuniformity of the grid spacing. After each smoothing sweep, the endpoint values are reset to the values of their interior neighbors.

$$P_1 = P_2$$

$$P_{i_{\max}} = P_{i_{\max}-1}$$

At least one smoothing step is recommended; however, it should be noted that repeated smoothing will reduce the function P defined in (1), and hence w , to a constant which would result in a uniform grid. The smoothing formula used here has worked better than the usual averaging techniques since it takes into consideration the grid distribution.

The weight function defined in (2) is for rezoning in the x -direction and will henceforth be denoted as w_x . The value of the weight function at $\xi = i$ is indicated by $w_{x,i}$. If x is replaced by y and i is replaced by j , an analogous weight function w_y can be derived for rezoning in the y -direction. Now the maximum weight function value in both directions can be computed as

$$w_{\max} = \max \left\{ \max_i w_{x,i}, \max_j w_{y,j} \right\}. \quad (4)$$

A parameter $\omega > 1$ will be introduced which will determine the degree to which the grid adapts to the weight functions. In order to see how this is done, consider the grid function

$$W_i = 1 + (\omega - 1) \frac{w_{x,i}}{w_{\max}}. \quad (5)$$

We are going to rezone so that the intervals Δx_i for the new grid will satisfy

$$\Delta x_i W_i = C_x \quad (6)$$

where C_x is a constant to be determined. Now it can be seen that if w_x is near the maximum value w_{\max} , then W_i is approximately equal to ω and $\Delta x_i \approx C_x/\omega$. On the other hand, if w_x is much less than w_{\max} , then W_i is nearly 1 and $\Delta x_i \approx C_x$. Therefore, when w_x assumes values near both extremes, the quantity ω is approximately the ratio of the maximum to minimum grid spacing. To determine the constant C_x , note that

$$x_{i_{\max}} - x_0 = \sum_{i=1}^{i_{\max}} \Delta x_i = \sum_{i=1}^{i_{\max}} \frac{C_x}{W_i} = C_x \sum_{i=1}^{i_{\max}} \frac{1}{W_i}$$

which implies that

$$C_x = (x_{i_{\max}} - x_0) \left[\sum_{i=1}^{i_{\max}} \frac{1}{W_i} \right]^{-1}. \quad (7)$$

The new grid intervals along the x -axis can be computed from (6) as

$$\Delta x_i = \frac{C_x}{W_i}$$

with C_x given in (7) and W_i given in (5), and the coordinates along the axis given as

$$x_i = \sum_{k=1}^i \Delta x_k.$$

A new distribution of grid points along the y -axis can also be computed using the following analogous formulas.

$$y_j = \sum_{k=1}^j \Delta y_k.$$

$$\Delta y_j = \frac{C_y}{W_j}$$

$$C_y = (y_{j_{\max}} - y_0) \left[\sum_{j=1}^{j_{\max}} \frac{1}{W_j} \right]^{-1}.$$

$$W_j = 1 + (\omega - 1) \frac{w_{y,j}}{w_{\max}}.$$

3 COMPUTATIONAL RESULTS

The adaptive rezoning procedure was used in the solution of several different types of problems involving the modeling of explosions. In all cases the grid was dynamically coupled with the solution so that the grid moved at every tenth time step. The choice of ten time steps between updates of the grid was chosen mostly by experience. It is often enough so that the grid is able to keep up with the solution. It is possible to rezone at each time step, but that is not advisable for two reasons. First of all, the computation time increases significantly, and secondly, the rezone procedure tends to smooth the solution so that pressure peaks are excessively damped. The rezone procedure was most successful in improving the qualitative nature of the numerical solution. It greatly reduced the oscillations (or ringing) in the solution without the addition of artificial viscosity. The adaptive grid algorithm was implemented with $\omega = 5$ to give a ratio of five for the maximum to minimum grid spacing. This had the effect of reducing the grid spacing in the neighborhood of the shock by a factor of approximately three.

The gridding scheme was capable of producing extremely fine grids near the shock, but the HULL code documentation [2] recommended that the grid aspect ratio not exceed three. This ratio was exceeded on some of our computations with no noticeable loss of accuracy, but in some cases erroneous results were obtained when the aspect ratio was extremely

large. The weight function was computed with $c_0 = c_1 = 1$ and $c_2 = 0$, since it is doubtful that the second derivatives can be reliably approximated for these types of problems. Only one smoothing iteration was used.

The first example is the computation of the underwater explosion of an infinite cylindrical charge of TNT as depicted in Figure 1. This was solved as a one-dimensional problem with a 500 by 2 grid. Figures 2 and 3 may be used to compare the solutions at $t = 0.01$ seconds computed using a uniform and an adaptive grid. The location of every tenth cell is indicated along the top borders of the plots. The adaptive grid clearly generated a smoother and more realistic pressure profile. However, neither peak pressure was near the theoretical value in Cole [3]. It would have taken three times as many grid points with a uniform grid to achieve the same resolution near the shock wave as was achieved using the adaptive grid. The lower peak pressure with the adaptive grid is primarily due to the smoothing effect of the interpolation between grids.

The next example is the solution of an axisymmetric problem. A charge of TNT in the shape of a cylinder is detonated underwater. The initial stage of the solution, illustrated in the axial and radial coordinate system, appears in Figure 4. The solutions computed on the uniform and adaptive grids are compared at two different times. Figure 5 is a plot of the solution on the uniform grid at an early time when the shock wave is near the charge. Figure 6 is the solution on the adaptive grid at the same time. Note the high concentration of grid points near the charge that is needed at the early stage of the computations in order to keep oscillations from initiating and propagating. Figures 7 and 8 are plots of solutions on the uniform and adaptive grids at a later time when the shock wave is near the outer boundary of the computational region. Oscillations in the solution computed on the uniform grid can be observed especially near the axis of symmetry. The solution also fails to develop into a spherically symmetric solution as discussed by Cole [3]. On the other hand, the solution computed on the adaptive grid develops into a nearly spherical solution with very little indication of grid orientation effects.

The adaptive rezoning scheme has also been used in the computation of airblasts. A spherical charge of HMX explosive is detonated in the atmosphere above the soil surface as illustrated in Figure 9. The solutions on the uniform and adaptive grids before the blast front reaches the surface are plotted in Figures 10 and 11. The grid for this problem is sufficiently fine so that both solutions appear quite reasonable. However, with the uniform grid there is a noticeable perturbation in what should be circular contour lines near the axis of symmetry. Both solutions were continued and correctly modelled the reflection of the shock wave off of the soil surface. The solution on the uniform grid is plotted in Figure 12. The solution on the adaptive grid appears in Figure 13. Note the concentration of grid points at the surface. This is due to the large pressure gradient between the air and soil which did not allow the grid to follow the reflected shock.

The final example is the computation of an underground explosion. A spherical charge of nitromethane is buried below the surface of the soil as indicated in Figure 14. The computed solutions on a uniform and adaptive grid are plotted in Figures 15 and 16. The coarse grid effects of the uniform grid are clearly evident in the noncircular contour lines. Both of these solutions were continued, but neither grid was capable of correctly modeling the shock wave as it passed through the air/soil interface.

4 CONCLUSIONS

It has been demonstrated that adaptive rezoning can significantly improve the quality of numerical solutions computed on a Cartesian grid system. Grid orientation effects were noticeably reduced in the examples included in this report. For problems involving explosions, the adaptive grid algorithm can be automated so that the grid points are concentrated in regions of high pressure gradient.

It should be noted that the capability of rezoning is limited by the fact that the HULL code uses only Cartesian grids. The ability to refine the grid locally would be a convenient feature in adapting the grid to the solution, but that would require a major restructuring of the code.

ACKNOWLEDGMENT

This work was sponsored by the Explosion Effects Division, Structures Laboratory, U S Army Corps of Engineers Waterways Experiment Station. Partial support was provided under the Battelle Scientific Services Agreement, Delivery Order 1308, Contract No. DAAL03-86-D-0001. Additional support was provided by Purchase Orders DACA39-90-M-3694, DACA39-91-M-1959, and DACA39-91-M-4034.

REFERENCES

1. R. D. Russell and J. Christiansen, "Adaptive Mesh Selection Strategies for Solving Boundary Value Problems," *SIAM Journal on Numerical Analysis*, 15 (1978) 59-80.
2. D. A. Matuska and J. J. Osborn, "HULL Documentation", Orlando Technology, Inc., Shalimar, Florida, 1987.
3. R. H. Cole, *Underwater Explosions*, Dover Publications, New York, 1965.

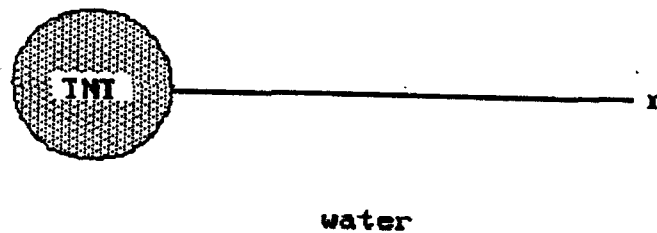


Figure 1. Initial configuration for an underwater explosion of an infinite cylindrical charge of TNT.

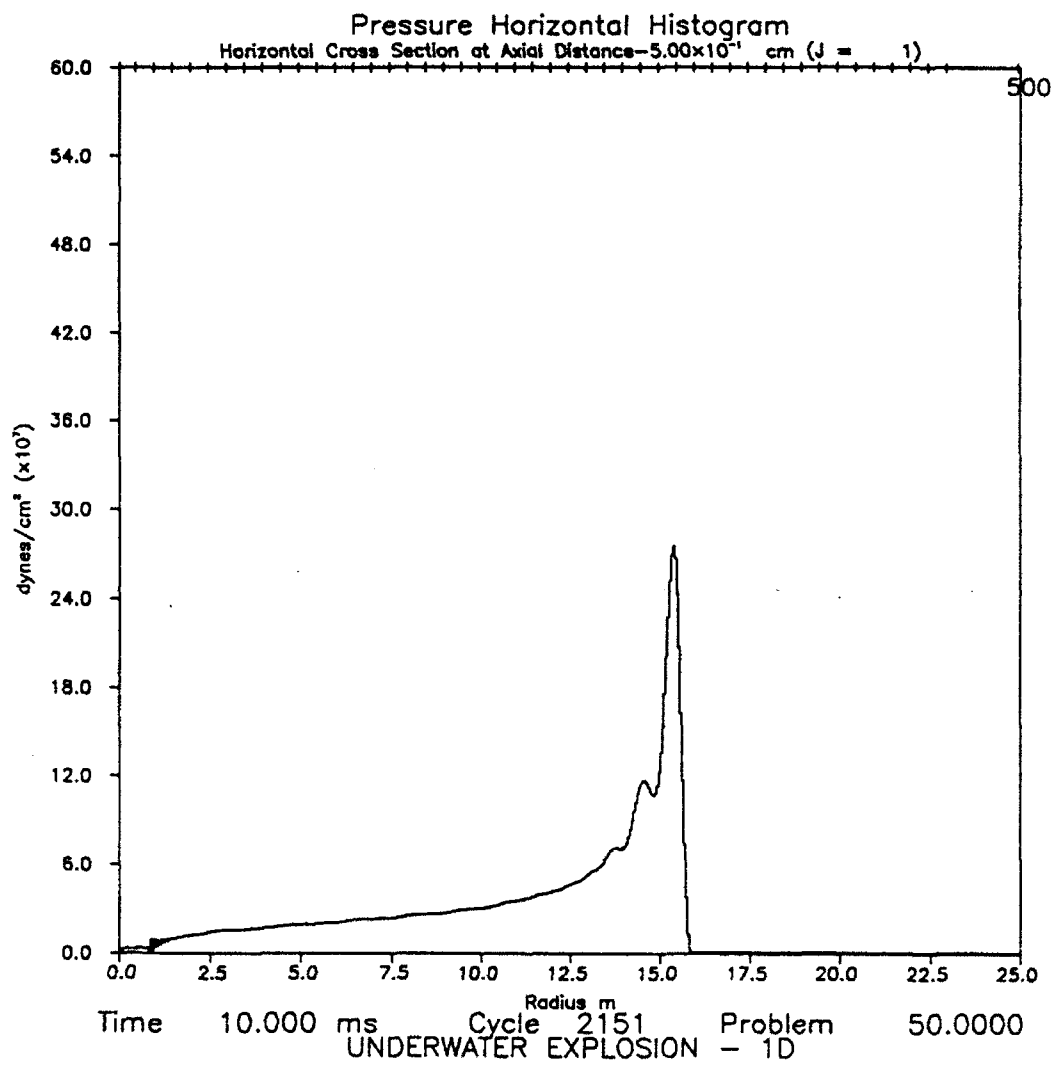


Figure 2. Pressure values calculated on a uniform grid at $t=10$ milliseconds.

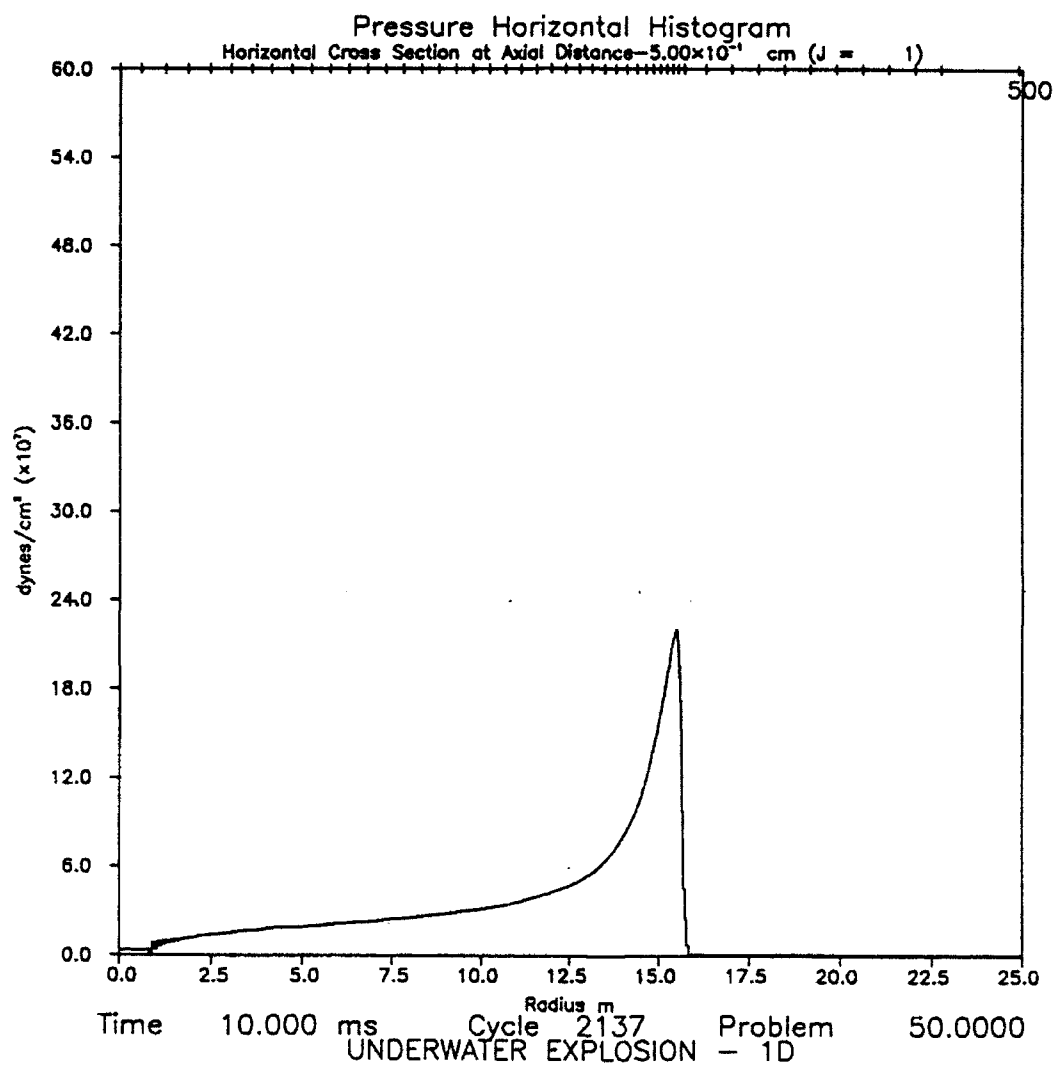


Figure 3. Pressure values calculated on an adaptive grid at $t=10$ milliseconds.

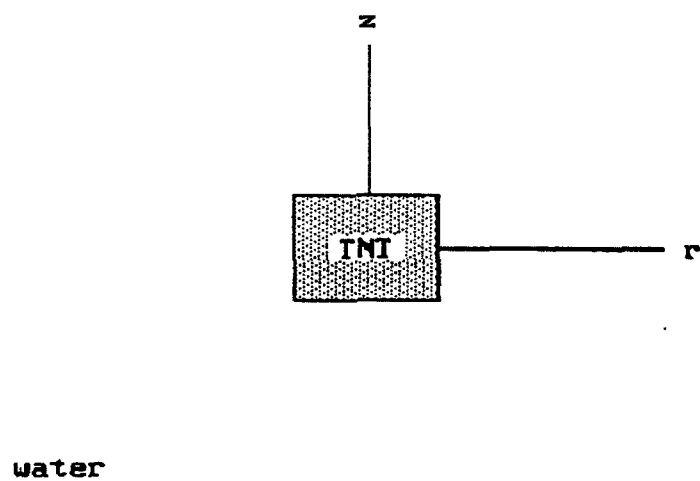


Figure 4. Initial configuration for an underwater explosion of a finite cylindrical charge of TNT.

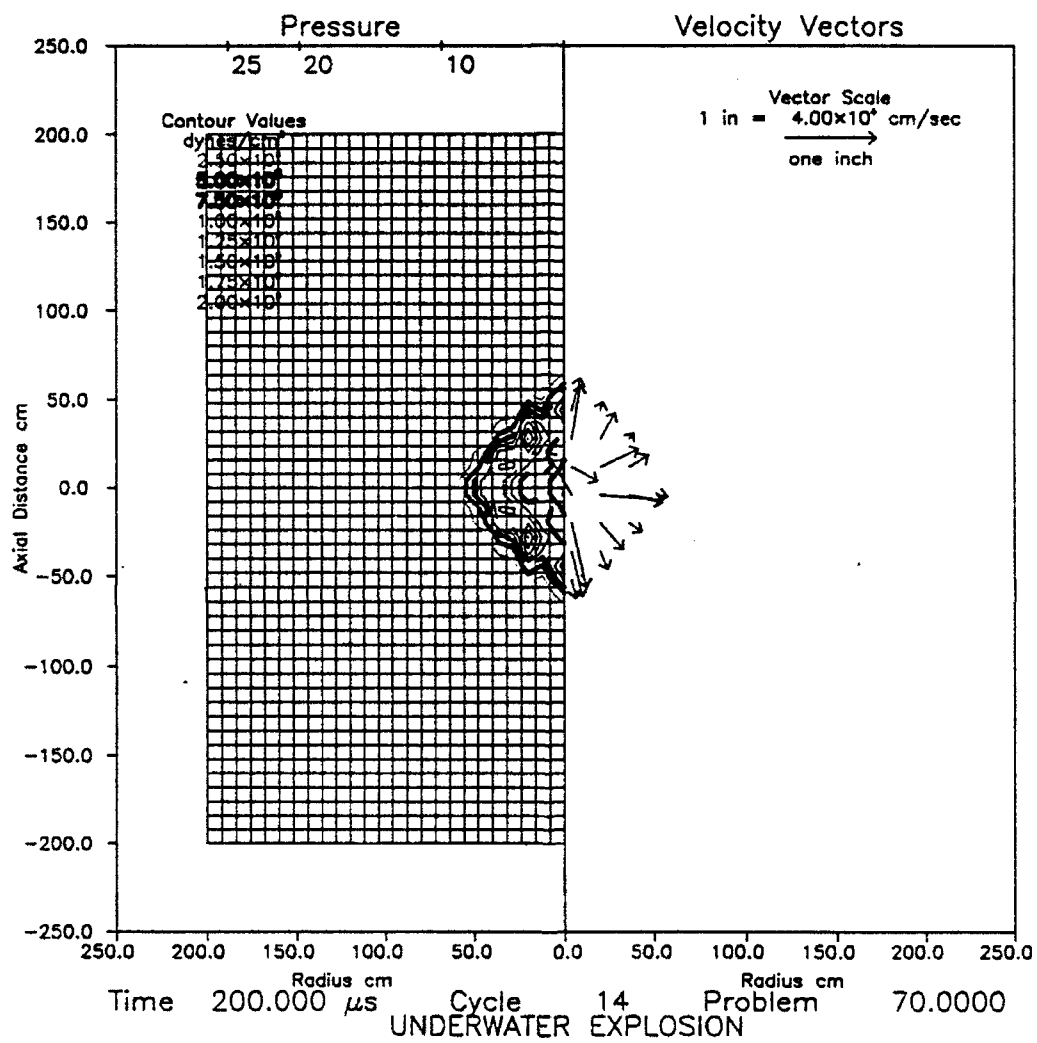


Figure 5. Pressure contours and velocity vectors calculated on a uniform grid at $t=200$ microseconds.

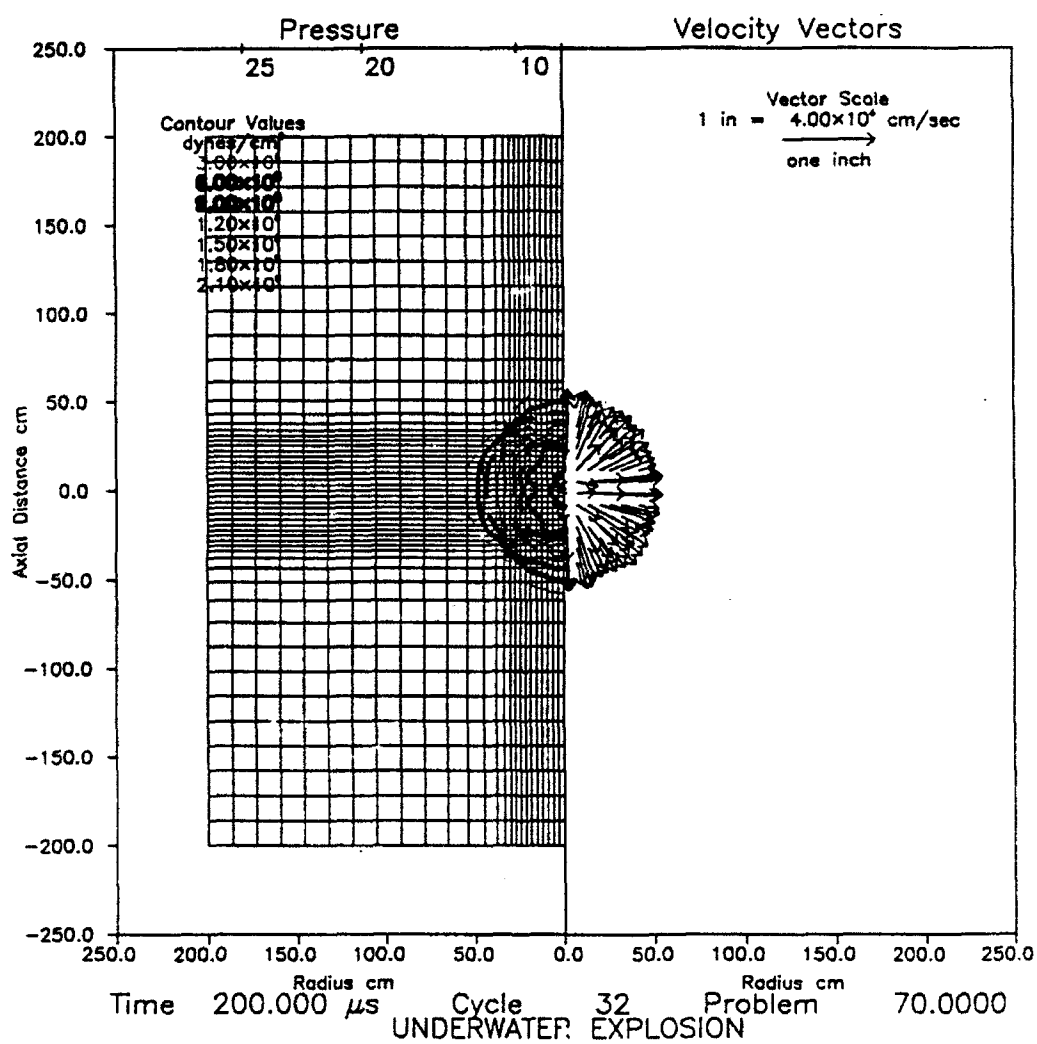


Figure 6. Pressure contours and velocity vectors calculated on an adaptive grid at $t=200$ microseconds.

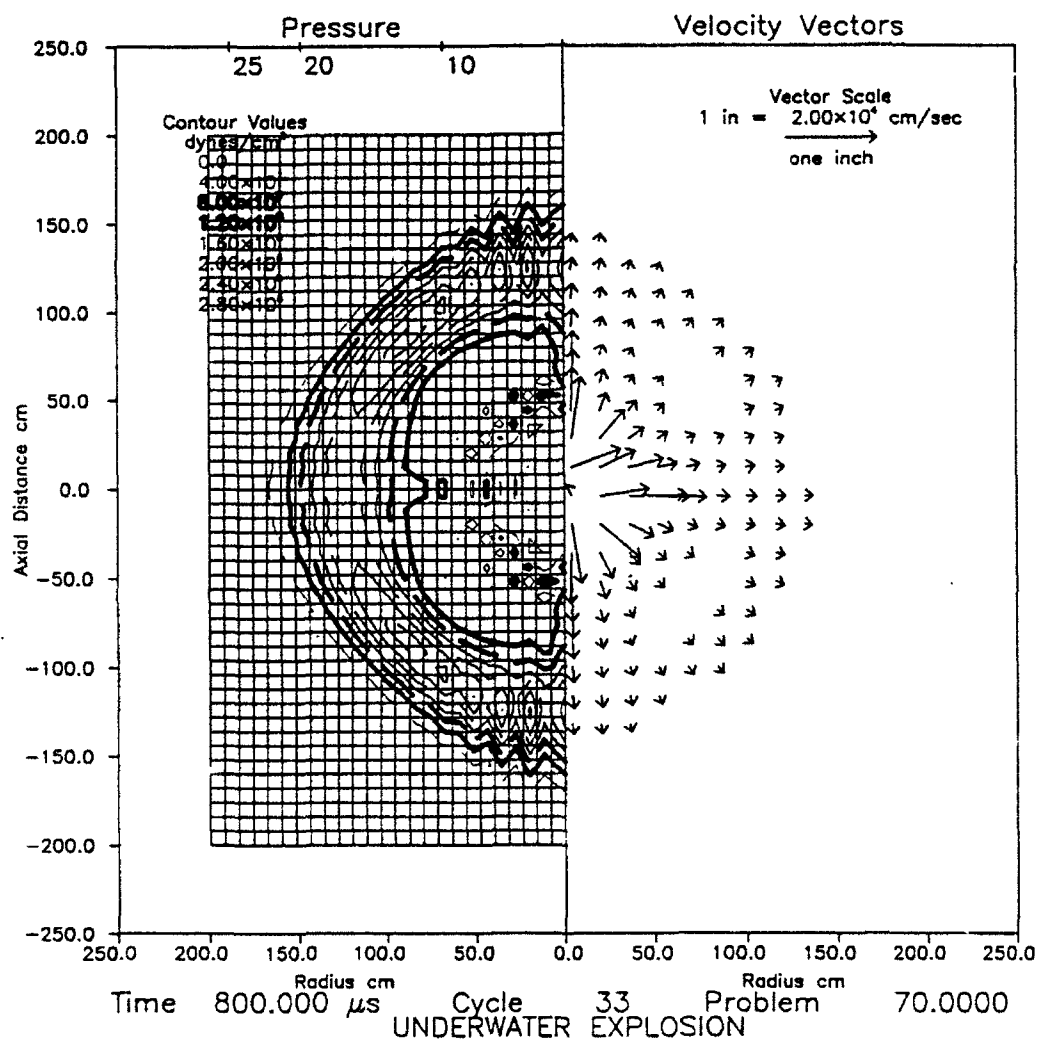


Figure 7. Pressure contours and velocity vectors calculated on a uniform grid at $t=200$ microseconds.

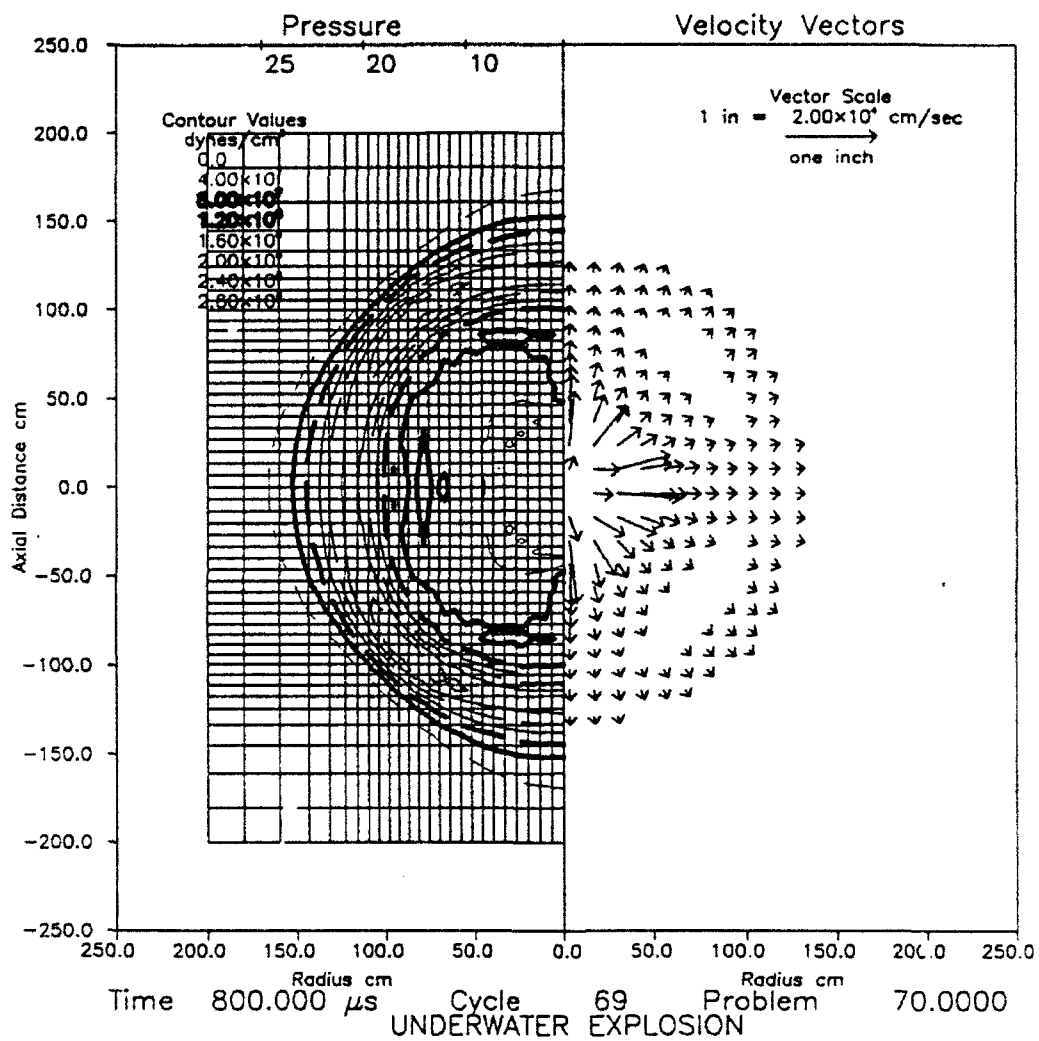


Figure 8. Pressure contours and velocity vectors calculated on an adaptive grid at $t=800$ microseconds.

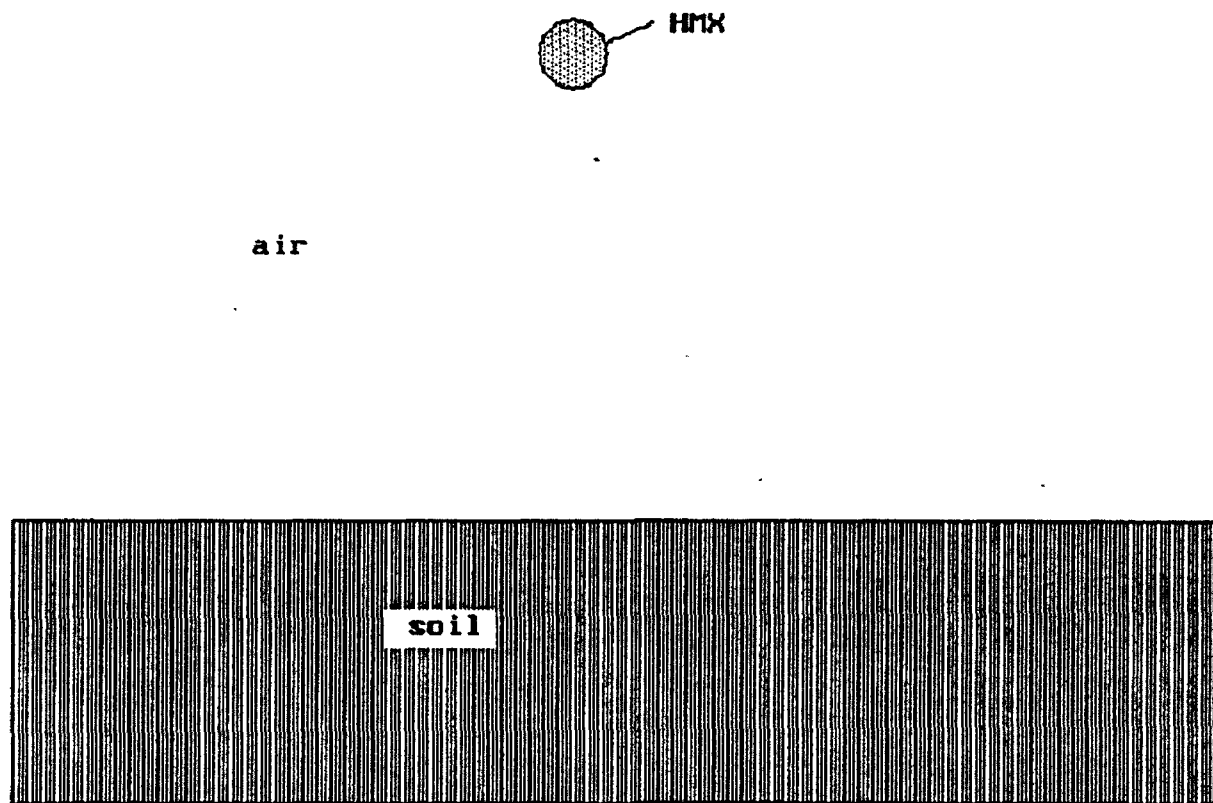


Figure 9. Initial configuration for an atmospheric explosion of a spherical charge of HMX.

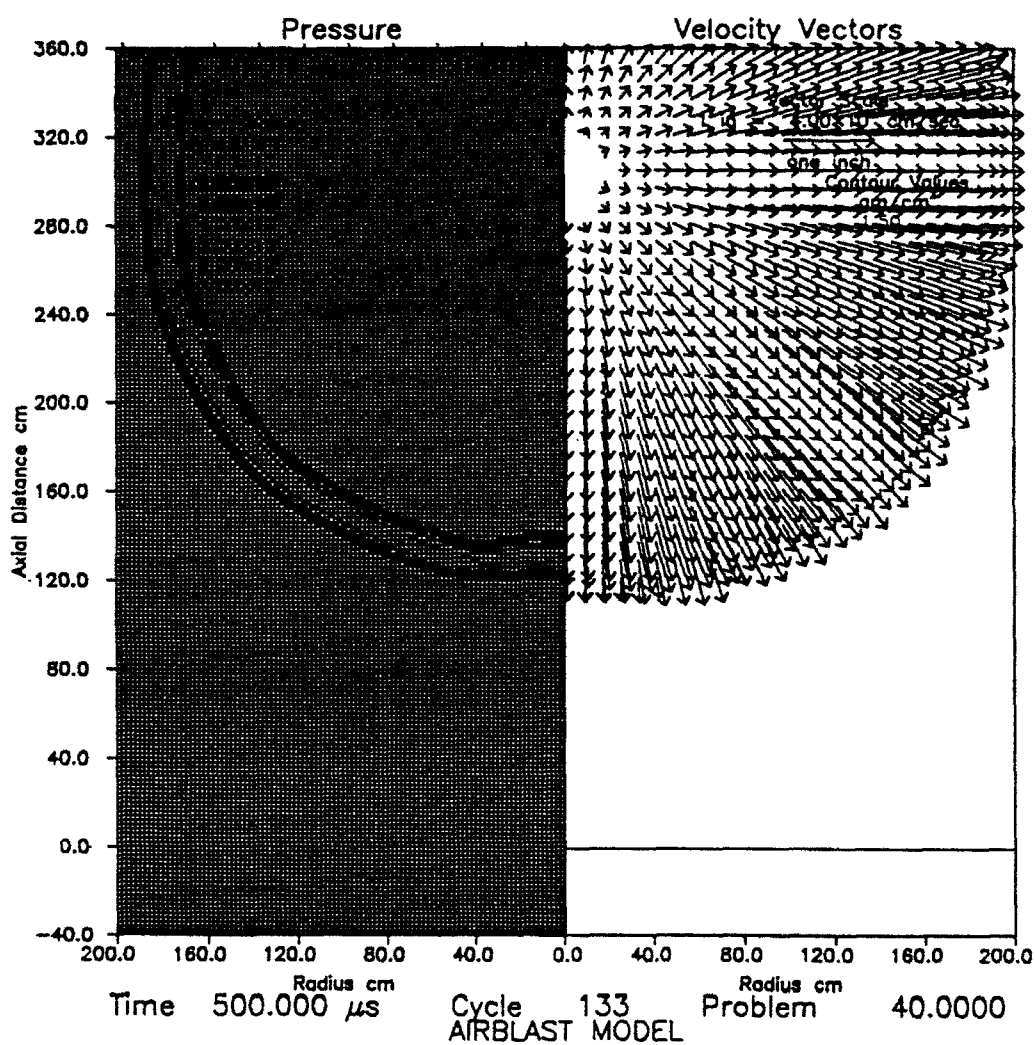


Figure 10. Pressure contours and velocity vectors calculated on a uniform grid at $t=500$ microseconds.

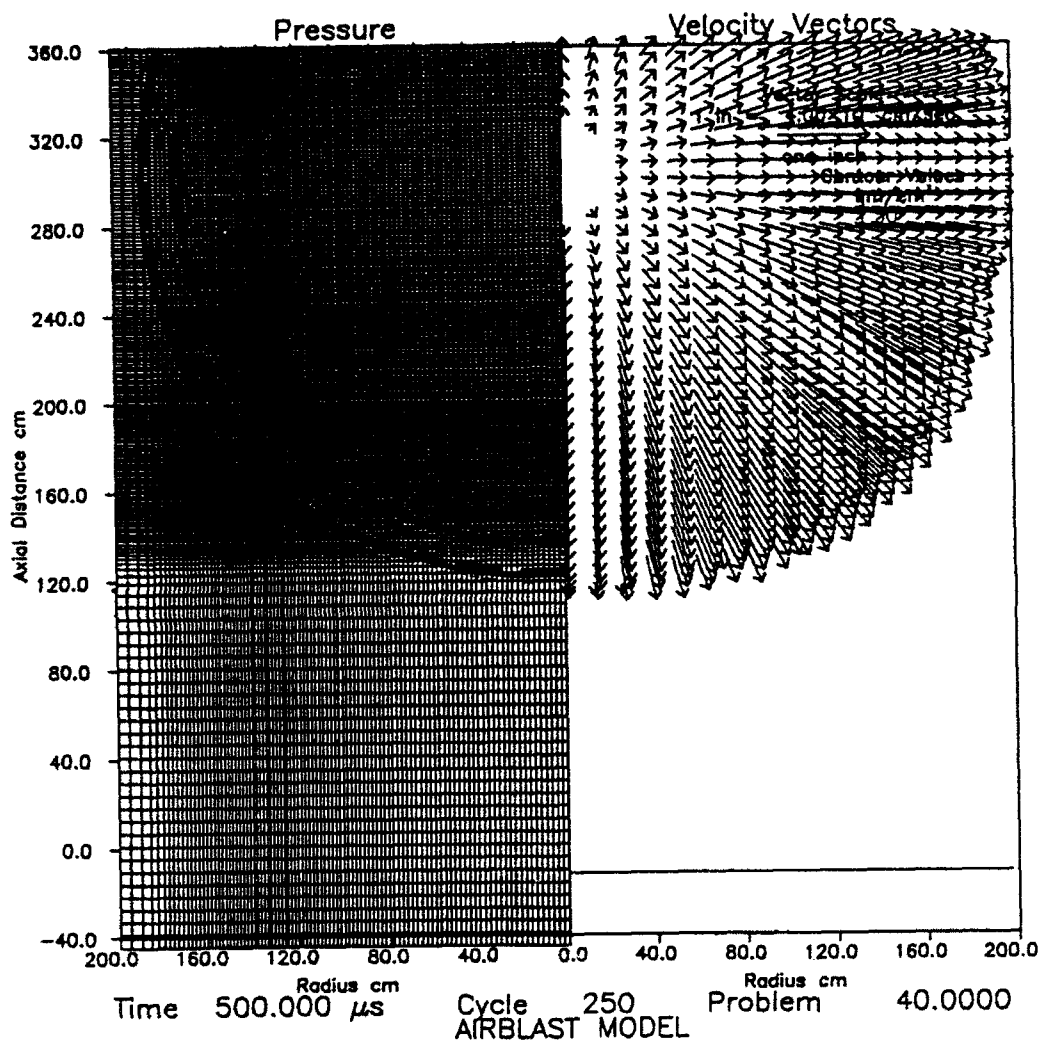


Figure 11. Pressure contours and velocity vectors calculated on an adaptive grid at $t=500$ microseconds.

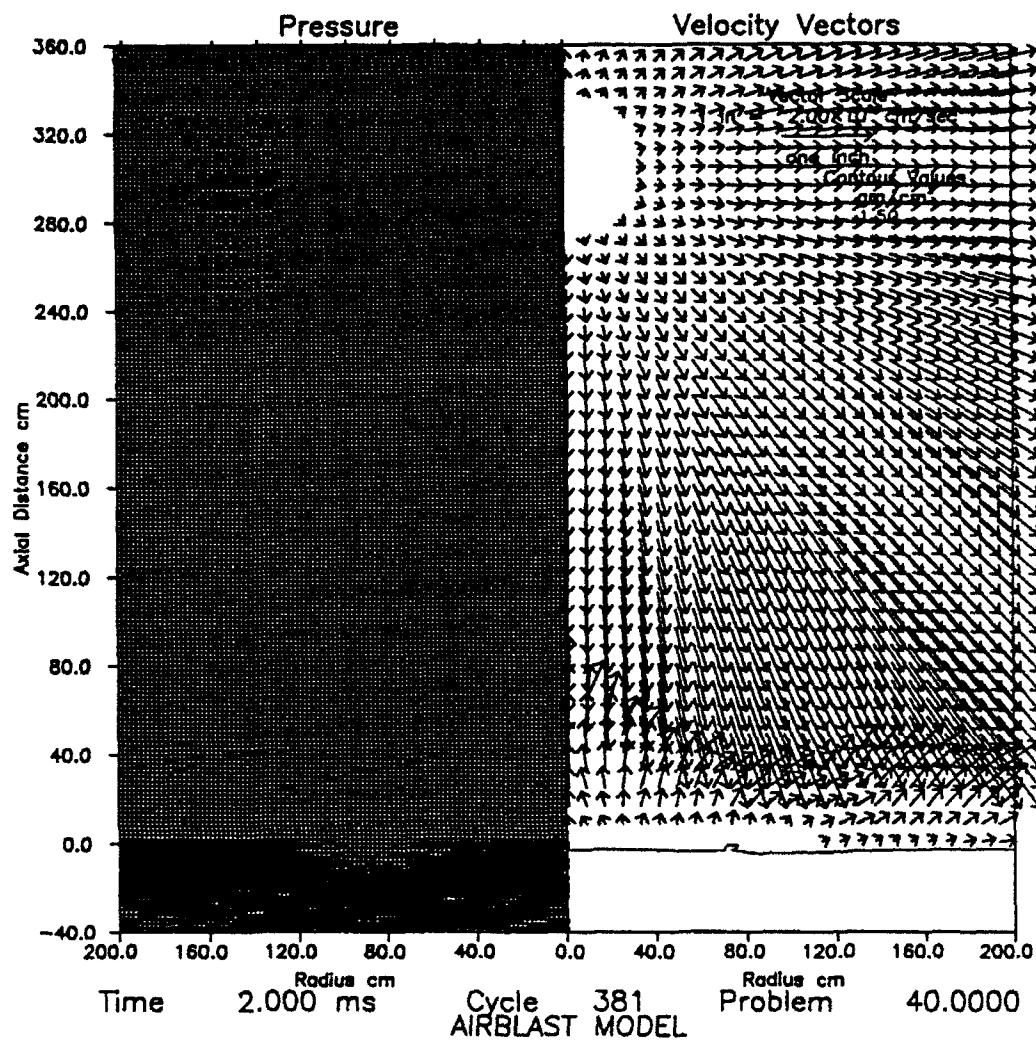


Figure 12. Pressure contours and velocity vectors calculated on a uniform grid at $t=2$ milliseconds.

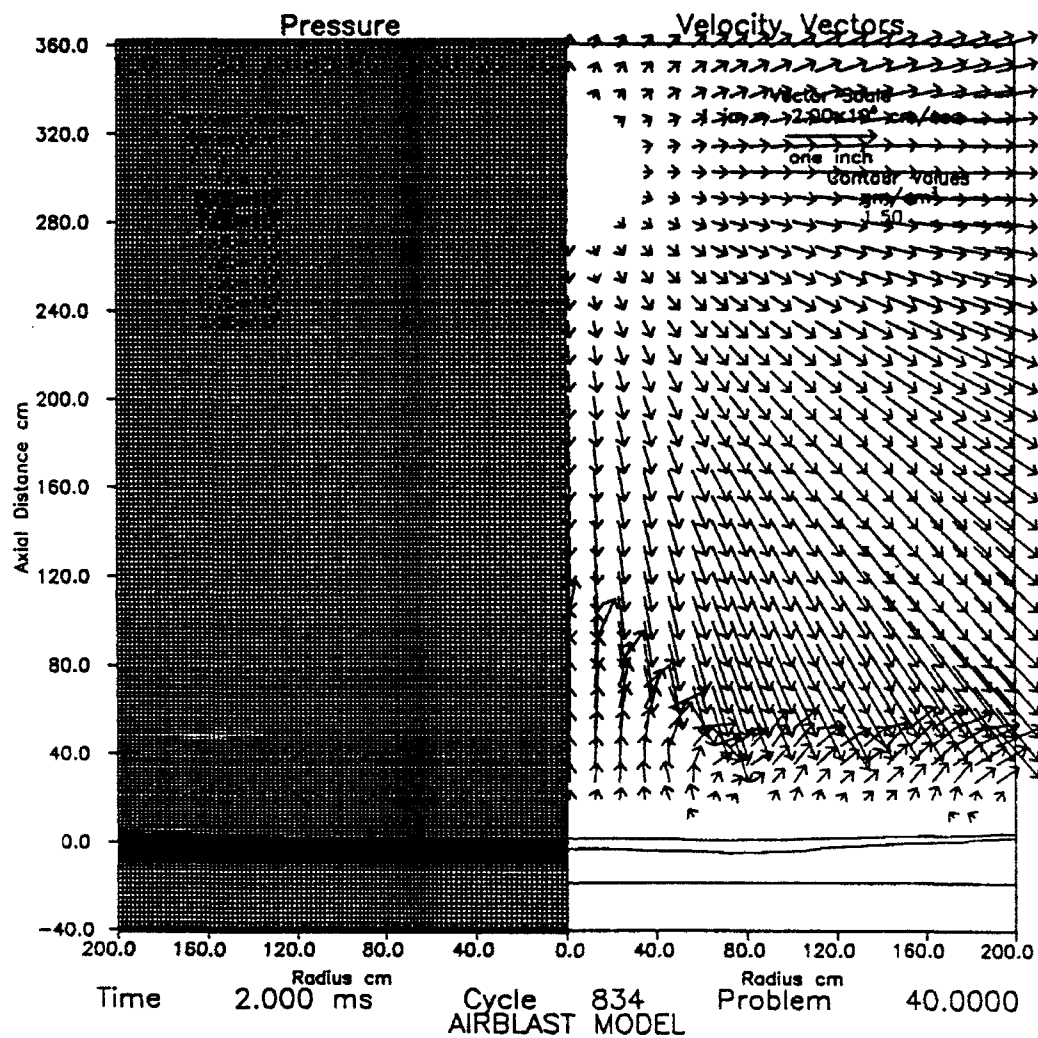


Figure 13. Pressure contours and velocity vectors calculated on an adaptive grid at $t=2$ milliseconds.

air

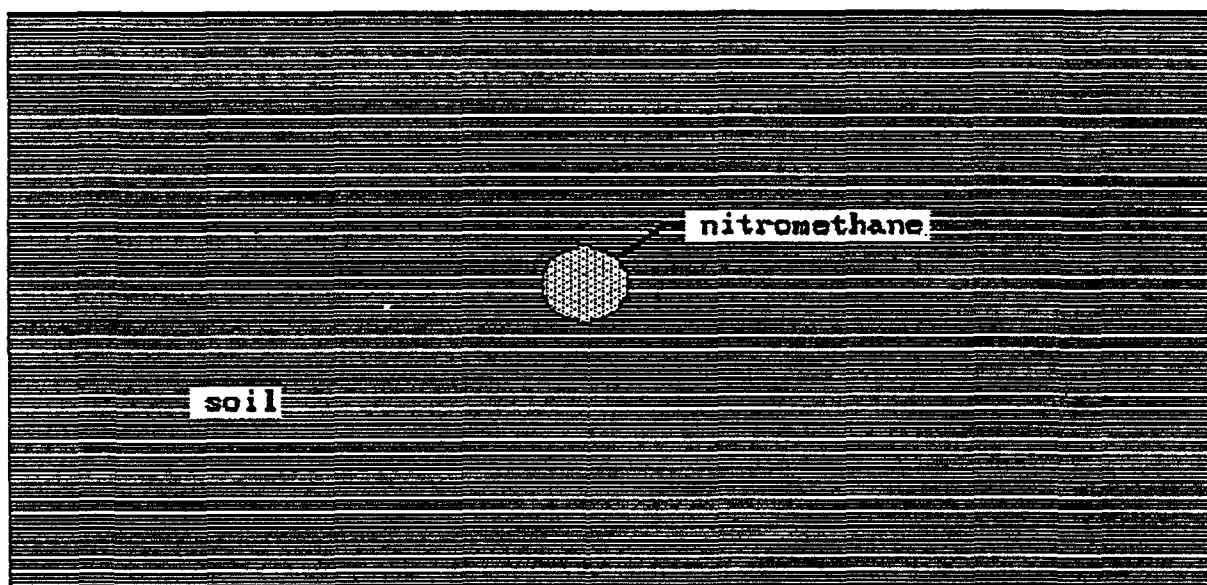


Figure 14. Initial configuration for an underground explosion of a spherical charge of nitromethane.

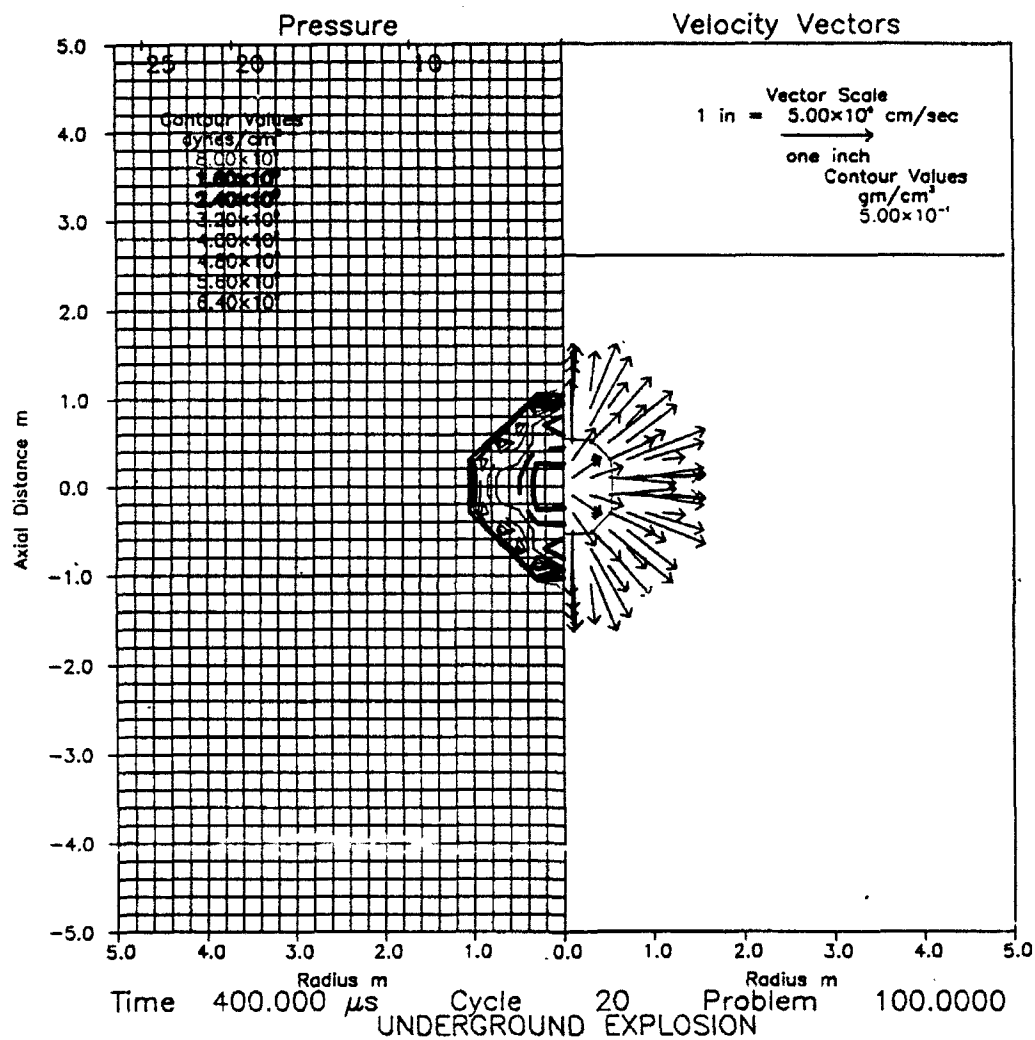


Figure 15. Pressure contours and velocity vectors calculated on a uniform grid at $t=400$ microseconds.

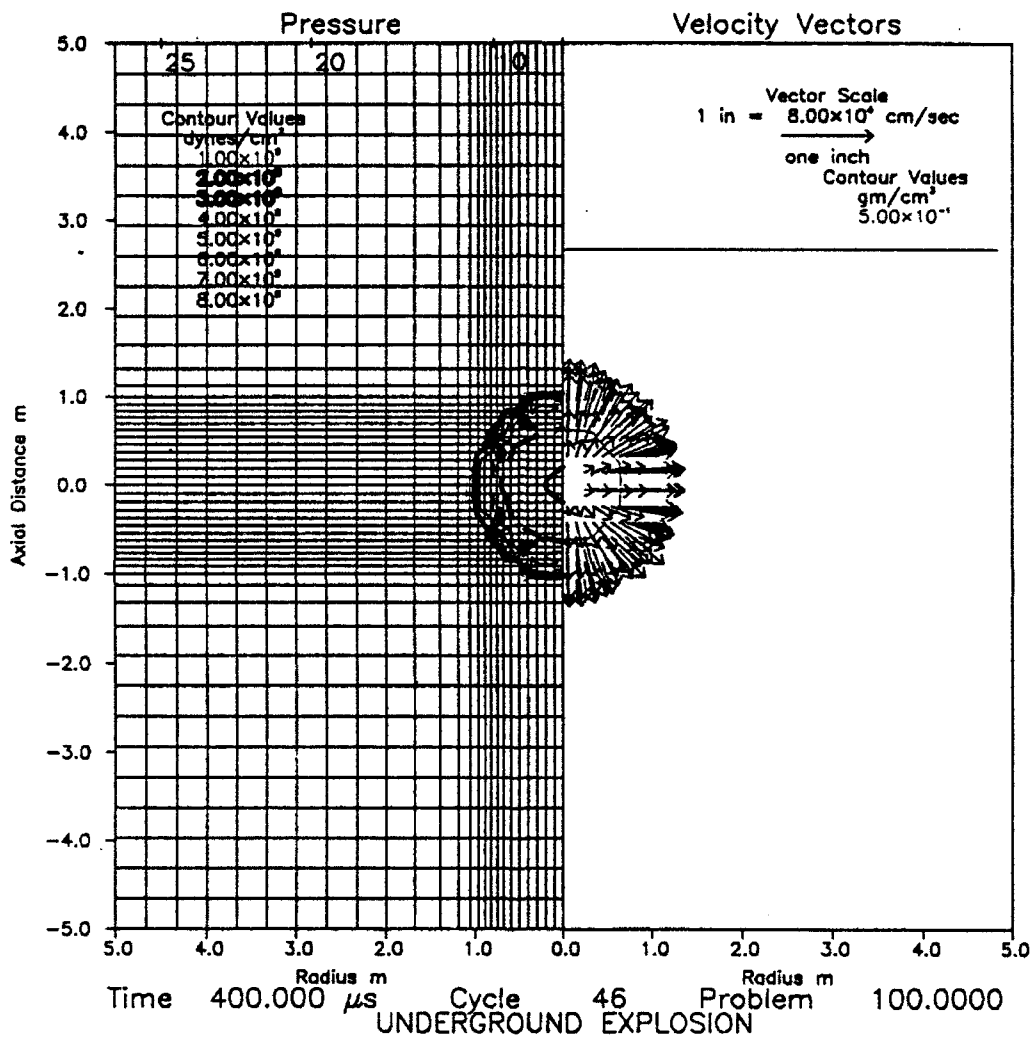


Figure 16. Pressure contours and velocity vectors calculated on an adaptive grid at $t=400$ microseconds.

FINITE DIMENSIONAL ESTIMATION ALGEBRAS OF MAXIMAL RANK WITH DIMENSION OF STATE SPACE EQUAL TO 3 *

Stephen S.-T. Yau, Jie Chen and Chi-Wah Leung

Department of Mathematics, Statistics
and Computer Science
University of Illinois at Chicago
Box 4348, M/C 249
Chicago, IL 60680

Abstract

The idea of using estimation algebras to construct finite dimensional nonlinear filters was first proposed in Brockett and Clark [1], Brockett [2] and Mitter [3]. In his famous talk at the International Congress of Mathematics in 1983, Brockett proposed to classify all finite dimensional estimation algebras. An affirmative solution to Brockett's problem will allow us to construct all possible finite dimensional recursive filters from the Lie algebraic point of view.

In 1990, the first named author [21] considered a general class of nonlinear filtering systems which include both Kalman-Bucy filtering systems and Benes filtering systems as special cases. A simple algebraic necessary and sufficient condition was established for an estimation algebra of this class of filtering systems to be finite dimensional. Consequently he has rigorously constructed a new class of finite dimensional filters which include both Kalman-Bucy filters and Benes filters as special cases. Note that the method used in [21] computes the fundamental solution of the D-M-Z equation and hence it also solves filtering problem with non-Gaussian initial conditions. In [5,22], the concept of an estimation algebra with maximal rank was introduced. This is the most important general subclass of estimation algebras because there is no assumption on the drift term of the nonlinear filtering system. The first named author and Chiou have already classified all maximal rank finite dimensional estimation algebra with state space dimension at most 2 [5,22]. In this report we continue the project and study the case for state space dimension 3. Consequently, we have shown that at least for low dimensional state space the finite dimensional filters constructed in [21] are the most general filters from Lie algebraic point of view.

1. Problem formulation

The filtering problem considered here is based on the following signal observation model:

$$(1.1) \quad \begin{cases} dx(t) = f(x(t))dt + g(x(t))dv(t) & x(0)=x_0 \\ dy(t) = h(x(t))dt + dw(t) & y(0)=y_0 \end{cases}$$

in which x, v, y and w are respectively $\mathbf{R}^n, \mathbf{R}^p, \mathbf{R}^m$ and \mathbf{R}^m valued processes, and v and w have components which are independent, standard Brownian processes. We further assume that $n = p, f, h$ are C^∞ smooth, and that g is an orthogonal matrix. We will refer to $x(t)$ as the state of the system at time t and to $y(t)$ as the observation at time t .

Let $\rho(t, x)$ denote the conditional density of the state given the observation $\{y(s) : 0 \leq s \leq t\}$. It is well known (see [8], for example) that $\rho(t, x)$ is given by normalizing a function, $\sigma(t, x)$, which satisfies the following Duncan-Mortensen-Zakai equation:

$$(1.2) \quad d\sigma(t, x) = L_0\sigma(t, x)dt + \sum_{i=1}^m L_i\sigma(t, x)dy_i(t), \quad \sigma(0, x) = \sigma_0$$

where

$$L_0 = \frac{1}{2} \sum_{i=1}^n \frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^n f_i \frac{\partial}{\partial x_i} - \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} - \frac{1}{2} \sum_{i=1}^m h_i^2$$

* Supported by the U.S. Army Research Office

and for $i = 1, \dots, m$, L_i is the zero order differential operator of multiplication by h_i . σ_0 is the probability density of the initial point x_0 . In this paper, we will assume σ_0 is a C^∞ function.

Definition : If X and Y are differential operators, the Lie bracket of X and Y , $[X, Y]$, is defined by

$$[X, Y]\psi = X(Y\psi) - Y(X\psi)$$

for any C^∞ function ψ .

Definition : The estimation algebra E , of a filtering problem (1.1) is defined to be the Lie algebra generated by $\{L_0, L_1, \dots, L_m\}$ or $E = \langle L_0, L_1, \dots, L_m \rangle_{L.A.}$. If in addition there exists a potential function ψ such that $f_i = \frac{\partial \psi}{\partial x_i}$ for all $1 \leq i \leq n$, then the estimation algebra is called exact.

Define

$$D_i = \frac{\partial}{\partial x_i} - f_i, \quad \text{and} \quad \eta = \sum_{i=1}^n \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^n f_i^2 + \sum_{i=1}^m h_i^2.$$

Then

$$L_0 = \frac{1}{2} \left(\sum_{i=1}^n D_i^2 - \eta \right)$$

We need the following basic results for later discussion.

Theorem 1. (Ocone) Let E be a finite dimensional estimation algebra. If a function $\xi \in E$, then ξ must be at most a quadratic polynomial.

Ocone's theorem ([12], see [6] for an extension) says that h_1, \dots, h_m in a finite dimensional estimation algebras are polynomials of degree ≤ 2 .

Definition : The estimation algebra E , of a filtering problem (1.1), is said to be the estimation algebra with maximal rank if $x_i + c_i \in E$ for all $1 \leq i \leq n$ where c_i is a constant.

Remark : If $m \geq n$, $h_i(x) = \sum_{j=1}^n a_{ij}x_j + c_i$ for $1 \leq i \leq m$ with $A = (a_{ij})_{n \times n}$ invertible, then E is of maximal rank.

The following theorem in [20] plays a fundamental role in the classification of all finite dimensional estimation algebras.

Theorem 2. Let E be a finite dimensional estimation algebra of (1.1) satisfying $\frac{\partial f_i}{\partial x_j} - \frac{\partial f_j}{\partial x_i} = \omega_{ij}$ where ω_{ij} are constants for all $1 \leq i \leq n$. Then h_1, \dots, h_m are polynomials of degree at most one. Furthermore, if E is of maximal rank, then η is a polynomial of degree at most 2 and E is a real Lie algebra spanned by $\{1, L_0, x_i, D_i : 1 \leq i \leq n\}$.

2. Classification of finite dimensional maximal rank estimation algebra

§2.1 Classification Theorem

Main Theorem : Suppose that the state space of the filtering system (1.1) is of dimension three. If E is the finite dimensional estimation algebra with maximal rank, then E is a real Lie algebra of dimension 8 with basis given by $\{1, x_1, x_2, x_3, D_1, D_2, D_3, L_0\}$.

In light of Theorem 2 above, it suffices for us to establish that $\omega_{ij} = \text{constant } \forall i, j$. Since $\omega_{ij} = [D_j, D_i]$ must be in E , ω_{ij} is a polynomial of degree at most two.

First of all, we'll show that ω_{ij} 's are polynomials of degrees at most one.

Let $w_{ij} = \text{degree 2 homogeneous part of } \omega_{ij}$

$$w_{12} = a_{12}x_1^2 + b_{12}x_2^2 + c_{12}x_3^2 + d_{12}x_1x_2 + e_{12}x_1x_3 + f_{12}x_2x_3$$

$$w_{13} = a_{13}x_1^2 + b_{13}x_2^2 + c_{13}x_3^2 + d_{13}x_1x_2 + e_{13}x_1x_3 + f_{13}x_2x_3$$

$$w_{23} = a_{23}x_1^2 + b_{23}x_2^2 + c_{23}x_3^2 + d_{23}x_1x_2 + e_{23}x_1x_3 + f_{23}x_2x_3$$

and

$$\begin{aligned}\frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2} &= \omega_{12} = w_{12} + p_{12} \\ \frac{\partial f_3}{\partial x_1} - \frac{\partial f_1}{\partial x_3} &= \omega_{13} = w_{13} + p_{13} \\ \frac{\partial f_3}{\partial x_2} - \frac{\partial f_2}{\partial x_3} &= \omega_{23} = w_{23} + p_{23}\end{aligned}$$

where p_{ij} is the linear part of ω_{ij} . Then we have the following basic cyclic relationship which is also true for general state dimension $n \geq 3$:

$$(2.1) \quad \frac{\partial \omega_{12}}{\partial x_3} + \frac{\partial \omega_{23}}{\partial x_1} + \frac{\partial \omega_{31}}{\partial x_2} = 0$$

Lemma 1: $W = (w_{ij})_{3 \times 3} = (0)_{3 \times 3}$ if any two of the w_{ij} 's with $i \neq j$ are 0.

Proof: There are $\binom{3}{2} = 3$ cases: (i) $w_{13} = w_{23} = 0$, (ii) $w_{23} = w_{12} = 0$, (iii) $w_{12} = w_{13} = 0$. Writing

$$\Omega = \begin{pmatrix} 0 & \omega_{12} & \omega_{13} \\ -\omega_{12} & 0 & \omega_{23} \\ -\omega_{13} & -\omega_{23} & 0 \end{pmatrix} \quad \text{and} \quad \frac{\partial}{\partial x} = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_3} \end{pmatrix},$$

then $\Omega \frac{\partial}{\partial x}(\omega_{kl})$ is a vector with entries in E (by considering $[[L_0, D_i], \omega_{kl}]$). For case (i), it suffices for us to consider operators $\omega_{12} \frac{\partial}{\partial x_1}$ and $\omega_{12} \frac{\partial}{\partial x_2}$ only. Operate them on ω_{12} and using Ocone's result, we conclude that W does not depend on x_1, x_2 i.e. $w_{12} = c_{12}x_3^2, w_{13} = w_{23} = 0$. The above cyclic relationship (2.1) implies that $2c_{12}x_3$ is a constant, hence c_{12} must be zero. Cases (ii) and (iii) can be treated similarly and we are done, i.e. $W = (0)_{3 \times 3}$ as desired.

Q.E.D.

Lemma 2: If one of the w_{ij} 's is zero where $i \neq j$, we have the following relationships among the coefficients in w_{ij} 's:

$$\begin{aligned}\text{(i)} \quad w_{23} = 0 &\Rightarrow \frac{\partial w_{13}}{\partial x_2} - \frac{\partial w_{12}}{\partial x_3} = 0 \quad \text{and} \quad d_{13} = e_{12}, 2b_{13} = f_{12}, f_{13} = 2c_{12}; \\ \text{(ii)} \quad w_{12} = 0 &\Rightarrow \frac{\partial w_{23}}{\partial x_1} - \frac{\partial w_{13}}{\partial x_2} = 0 \quad \text{and} \quad d_{13} = 2a_{23}, 2b_{13} = d_{23}, f_{13} = e_{23}; \\ \text{(iii)} \quad w_{13} = 0 &\Rightarrow \frac{\partial w_{23}}{\partial x_1} + \frac{\partial w_{12}}{\partial x_3} = 0 \quad \text{and} \quad e_{12} = -2a_{23}, 2c_{12} = -e_{23}, f_{12} = -d_{23}.\end{aligned}$$

Proof: These are easily verified by equation (2.1).

Q.E.D.

§2.2 Ω is linear

For the proof of $W = (0)_{3 \times 3}$ in general, we define the operators:

$$\begin{aligned}D1 &:= \omega_{12} \frac{\partial}{\partial x_2} + \omega_{13} \frac{\partial}{\partial x_3} \\ D2 &:= -\omega_{12} \frac{\partial}{\partial x_1} + \omega_{23} \frac{\partial}{\partial x_3} \\ D3 &:= -\omega_{13} \frac{\partial}{\partial x_1} - \omega_{23} \frac{\partial}{\partial x_2}\end{aligned}$$

Since E contains $Y_j = [L_0, D_j]$, so $[Y_j, \omega_{kl}] = \sum_{i=1}^3 \omega_{ji} \frac{\partial}{\partial x_i}(\omega_{kl}) \in E$, i.e. $\sum_{i=1}^3 \omega_{ji} \frac{\partial}{\partial x_i}(\omega_{kl})$ are polynomials of degree at most two. We may set the coefficients of terms with degree 3 equal to zero. But these terms arise from degree 2 terms of ω_{ji} 's only, therefore it suffices for us to consider w_{ij} instead of ω_{ij} in the following calculation.

Let

$Diw_{jk}300 :=$ coefficient of x_1^3 of $Di(w_{jk})$
 $Diw_{jk}030 :=$ coefficient of x_2^3 of $Di(w_{jk})$
 $Diw_{jk}003 :=$ coefficient of x_3^3 of $Di(w_{jk})$
 $Diw_{jk}210 :=$ coefficient of $x_1^2x_2$ of $Di(w_{jk})$
 $Diw_{jk}201 :=$ coefficient of $x_1^2x_3$ of $Di(w_{jk})$
 $Diw_{jk}120 :=$ coefficient of $x_2^2x_1$ of $Di(w_{jk})$
 $Diw_{jk}021 :=$ coefficient of $x_2^2x_3$ of $Di(w_{jk})$
 $Diw_{jk}102 :=$ coefficient of $x_3^2x_1$ of $Di(w_{jk})$
 $Diw_{jk}012 :=$ coefficient of $x_3^2x_2$ of $Di(w_{jk})$
 $Diw_{jk}111 :=$ coefficient of $x_1x_2x_3$ of $Di(w_{jk})$

Objective : To show that all coefficients of w_{ij} 's are zero, i.e. the system of 90 equations has only trivial solution.

Observe that $D1w_{12}003 = c_{12}(f_{12} + 2c_{13})$.

Suppose $c_{12} \neq 0 \Rightarrow f_{12} = -2c_{13}$, $D1w_{13}030 = b_{13}(f_{13} + 2b_{12})$.

Suppose $b_{13} \neq 0 \Rightarrow f_{13} = -2b_{12} \Rightarrow D2w_{12}003 = c_{12}(2c_{23} - e_{12})$, $D3w_{13}030 = -b_{13}(d_{13} + 2b_{23})$.
 $\Rightarrow e_{12} = 2c_{23}$, $d_{13} = -2b_{23} \Rightarrow D2w_{23}300 = a_{23}(e_{23} - 2a_{12})$, $D3w_{23}300 = -a_{23}(d_{23} + 2a_{13})$.

Suppose $a_{23} \neq 0 \Rightarrow e_{23} = 2a_{12}$, $d_{23} = -2a_{13} \Rightarrow D1w_{23}030 = 2b_{12}b_{23} + b_{13}f_{23}$, $D3w_{12}030 = -2b_{12}b_{23} - b_{13}d_{12}$. Add these two gives $b_{13}(f_{23} - d_{12}) = 0$. Since $b_{13} \neq 0$ is assumed, we have $f_{23} = d_{12}$.

Suppose $d_{12} \neq 0 \Rightarrow b_{12}, b_{23}, c_{13}, c_{23}, a_{23}, a_{13} \neq 0$ since $D1w_{23}030 = 2b_{12}b_{23} + d_{12}b_{13}$, $D1w_{23}003 = c_{12}d_{12} + 2c_{23}c_{13}$, $D2w_{23}210 = -2a_{12}a_{13} - d_{12}a_{23}$. Moreover from $D1w_{12}021$, $D1w_{23}120$, $D2w_{12}201$, and $D2w_{12}021$ we can obtain the following proportional relationships: $a_{23}/a_{12} = b_{23}/b_{12} = c_{23}/c_{12}$, $a_{13}/a_{12} = b_{13}/b_{12} = c_{13}/c_{12}$ and $a_{23}/a_{13} = b_{23}/b_{13} = c_{23}/c_{13}$.

Observe that $w_{23} = a_{23}x_1^2 + b_{23}x_2^2 + c_{23}x_3^2 - 2a_{13}x_1x_2 + 2a_{12}x_1x_3 + d_{12}x_2x_3$. Using the expressions $-d_{12}/2 = a_{12}a_{13}/a_{23} = b_{12}b_{23}/b_{13} = c_{13}c_{23}/c_{12}$ it can be shown that $a_{23}w_{23} = (a_{23}x_1 - a_{13}x_2 + a_{12}x_3)^2$. Similarly, $a_{23}^2w_{12} = a_{12}(a_{23}x_1 - a_{13}x_2 + a_{12}x_3)^2$ and $a_{23}^2w_{13} = a_{13}(a_{23}x_1 - a_{13}x_2 + a_{12}x_3)^2$. The cyclic relation (2.1) implies that $0 = \frac{\partial w_{23}}{\partial x_1} + \frac{\partial w_{31}}{\partial x_2} + \frac{\partial w_{12}}{\partial x_3} = (a_{23}x_1 - a_{13}x_2 + a_{12}x_3)^2(1 + \frac{a_{12}^2}{a_{23}^2} + \frac{a_{13}^2}{a_{23}^2})$. Contradicting that $a_{12}, a_{13}, a_{23} \neq 0$. Hence $d_{12} \neq 0$ under $c_{12}, b_{13}, a_{23} \neq 0$ is impossible.

Then we are working with $d_{12} = 0$ under the assumptions that $c_{12}, b_{13}, a_{23} \neq 0 \Rightarrow D2w_{12}111 = -4a_{13}c_{12} \Rightarrow a_{13} = 0 \Rightarrow D1w_{23}120 = 2a_{12}b_{13}$, $D2w_{12}210 = -2a_{23}c_{13}$, $D2w_{13}300 = a_{23}e_{13}$, $D3w_{13}300 = -a_{23}d_{13}$, $D3w_{13}210 = -2a_{23}b_{13}$, $D3w_{23}030 = -2b_{23}^2 \Rightarrow a_{12} = c_{13} = e_{13} = d_{13} = b_{13} = b_{23} = 0 \Rightarrow D2w_{12}201 = 2a_{23}c_{12}$. Contradicting that $a_{23} \neq 0$ and $c_{12} \neq 0$. So it's impossible for $a_{23} \neq 0$ under $c_{12}, b_{13} \neq 0$.

Then $a_{23} = 0$ under the assumptions that $c_{12}, b_{13} \neq 0 \Rightarrow D2w_{12}300 = -2a_{12}^2$, $D3w_{13}300 = -2a_{13}^2 \Rightarrow a_{12} = 0, a_{13} = 0 \Rightarrow D2w_{23}201 = e_{23}^2$, $D3w_{23}210 = -d_{23}^2 \Rightarrow e_{23} = 0 \Rightarrow D2w_{23}003 = 2c_{23}^2$, $D3w_{23}030 = -2b_{23}^2 \Rightarrow c_{23} = b_{23} = 0 \Rightarrow D1w_{23}030 = b_{13}f_{23}$, $D2w_{12}003 = -c_{12}e_{12}$, $D2w_{12}120 = -d_{12}^2$, $D2w_{13}003 = -e_{13}c_{12}$, $D3w_{13}120 = -d_{13}^2 \Rightarrow f_{23} = e_{12} = d_{12} = e_{13} = d_{13} = 0$. Now $w_{23} = 0$. (2.1) gives $c_{13} = -b_{13}$. Then $D1w_{12}030 = 2b_{12}^2 - 2b_{13}c_{13} = 2(b_{12}^2 + b_{13}^2)$, so $b_{13} = 0$. Contradiction. So $b_{13} \neq 0$ cannot come together with $c_{12} \neq 0$.

Then $b_{13} = 0$ under the assumption $c_{12} \neq 0 \Rightarrow D1w_{12}030 = 2b_{12}^2$, $D3w_{23}030 = -2b_{23}^2 \Rightarrow b_{12} = 0$, $b_{23} = 0 \Rightarrow D1w_{13}021 = f_{13}^2$, $D3w_{13}120 = -d_{13}^2 \Rightarrow f_{13} = 0, d_{13} = 0 \Rightarrow D1w_{12}012 = 2c_{13}^2$, $D3w_{13}300 = -2a_{13}^2 \Rightarrow c_{13} = 0, a_{13} = 0 \Rightarrow D1w_{12}210 = d_{12}^2$, $D1w_{13}201 = e_{13}^2$, $D2w_{23}021 = f_{23}^2$, $D3w_{23}210 = -d_{23}^2 \Rightarrow d_{12} = e_{13} = f_{23} = d_{23} = 0 \Rightarrow w_{13} = 0$. By Lemma 2 $e_{23} = -2c_{12}$. Together with $D2w_{23}003 = -e_{23}c_{12} + 2c_{23}^2$ we obtain $0 = 2(c_{12}^2 + c_{23}^2)$. Contradicting $c_{12} \neq 0$. Then c_{12} cannot be non-zero.

Now we knew that $c_{12} = 0$ and no other restrictions. Then $D1w_{13}003 = 2c_{13}^2$, $D2w_{23}003 = 2c_{23}^2 \Rightarrow c_{13} = 0, c_{23} = 0 \Rightarrow D1w_{12}012 = f_{12}^2$, $D2w_{12}102 = -e_{12}^2 \Rightarrow f_{12} = 0, e_{12} = 0 \Rightarrow D1w_{12}030 = 2b_{12}^2$, $D2w_{12}300 = -2a_{12}^2 \Rightarrow b_{12} = a_{12} = 0 \Rightarrow D1w_{12}210 = d_{12}^2$, $D1w_{13}201 = e_{13}^2$, $D1w_{13}021 = f_{13}^2$, $D2w_{23}201 = e_{23}^2$, $D2w_{23}201 = f_{23}^2 \Rightarrow d_{12} = e_{13} = f_{13} = e_{23} = f_{23} = 0 \Rightarrow w_{12} = 0$. Lemma 2 gives $d_{13} = 2a_{23}$, $d_{23} = 2b_{13}$. Together with $D3w_{13}300 = -2a_{13}^2 - a_{23}d_{13}$ and $D3w_{23}030 = -d_{23}b_{13} - 2b_{23}^2$, we can conclude that $w_{13} = w_{23} = 0$ also.

$\Omega = (\omega_{ij})_{3 \times 3}$ is shown to be linear.

§2.2.1 Appendix

There are 90 quadratic equations arised from the coefficients :

$D_1 w_{12}$

300, 030, 003	$a_{12}d_{12} + a_{13}e_{12}, 2b_{12}^2 + b_{13}f_{12}, c_{12}f_{12} + 2c_{13}c_{12}$
210	$d_{12}^2 + 2a_{12}b_{12} + a_{13}f_{12} + d_{13}e_{12}$
102	$a_{12}f_{12} + e_{12}d_{12} + 2a_{13}c_{12} + e_{13}e_{12}$
120	$3b_{12}d_{12} + b_{13}e_{12} + d_{13}f_{12}$
021	$3b_{12}f_{12} + 2b_{13}c_{12} + f_{13}f_{12}$
102	$c_{12}d_{12} + e_{12}f_{12} + c_{13}e_{12} + 2e_{13}c_{12}$
012	$f_{12}^2 + 2c_{12}b_{12} + c_{13}f_{12} + 2f_{13}c_{12}$
111	$2d_{12}f_{12} + 2e_{12}b_{12} + 2d_{13}c_{12} + e_{13}f_{12} + f_{13}e_{12}$

$D_1 w_{13}$

300, 030, 003	$a_{12}d_{13} + a_{13}e_{13}, 2b_{12}b_{13} + b_{13}f_{13}, 2c_{13}^2 + f_{13}c_{12}$
210	$2a_{12}b_{13} + d_{12}d_{13} + a_{13}f_{13} + d_{13}e_{13}$
201	$a_{12}f_{13} + d_{13}e_{12} + 2a_{13}c_{13} + e_{13}^2$
120	$b_{12}d_{13} + 2d_{12}b_{13} + b_{13}e_{13} + d_{13}f_{13}$
021	$b_{12}f_{13} + 2b_{13}f_{12} + 2b_{13}c_{13} + f_{13}^2$
102	$d_{13}c_{12} + f_{13}e_{12} + 3c_{13}e_{13}$
021	$2b_{13}c_{12} + f_{13}f_{12} + 3c_{13}f_{13}$
111	$d_{12}f_{13} + 2b_{13}e_{12} + d_{13}f_{12} + 2d_{13}c_{13} + 2e_{13}f_{13}$

$D_1 w_{23}$

300, 030, 003	$a_{12}d_{23} + a_{13}e_{23}, 2b_{12}b_{23} + b_{13}f_{23}, c_{12}f_{23} + 2c_{13}c_{23}$
210	$2a_{12}b_{23} + d_{12}d_{23} + a_{13}f_{23} + d_{13}e_{23}$
201	$a_{12}f_{23} + e_{12}d_{23} + 2a_{13}c_{23} + e_{13}e_{23}$
120	$b_{12}d_{23} + 2d_{12}b_{23} + b_{13}e_{23} + d_{13}f_{23}$
021	$b_{12}f_{23} + 2f_{12}b_{23} + 2b_{13}c_{23} + f_{13}f_{23}$
102	$c_{12}d_{23} + e_{12}f_{23} + c_{13}e_{23} + 2e_{13}c_{23}$
012	$2c_{12}b_{23} + f_{12}f_{23} + c_{13}f_{23} + 2f_{13}c_{23}$
111	$d_{12}f_{23} + 2e_{12}b_{23} + f_{12}d_{23} + 2d_{13}c_{23} + e_{13}f_{23} + f_{13}e_{23}$

$D_2 w_{12}$

300, 030, 003	$-2a_{12}^2 + a_{23}e_{12}, -b_{12}d_{12} + f_{12}b_{23}, -c_{12}e_{12} + 2c_{23}c_{12}$
210	$-3a_{12}d_{12} + a_{23}f_{12} + e_{12}d_{23}$
201	$-3a_{12}e_{12} + 2a_{23}c_{12} + e_{23}e_{12}$
120	$-2a_{12}b_{12} - d_{12}^2 + e_{12}b_{23} + f_{12}d_{23}$
021	$-e_{12}b_{12} - d_{12}f_{12} + 2c_{12}b_{23} + f_{12}f_{23}$
102	$-2c_{12}a_{12} - e_{12}^2 + c_{23}e_{12} + 2e_{23}c_{12}$
012	$-c_{12}d_{12} - e_{12}f_{12} + c_{23}f_{12} + 2c_{12}f_{23}$
111	$-2e_{12}d_{12} - 2a_{12}f_{12} + 2c_{12}d_{23} + e_{23}f_{12} + e_{12}f_{23}$

$D_2 w_{13}$

300, 030, 003	$-2a_{12}a_{13} + a_{23}e_{13}, -b_{12}d_{13} + b_{23}f_{13}, 2c_{13}c_{23} - e_{13}c_{12}$
210	$-a_{12}d_{13} - 2d_{12}a_{13} + a_{23}f_{13} + d_{23}e_{13}$
201	$e_{13}e_{23} - a_{12}e_{13} - 2a_{13}e_{12} + 2a_{23}c_{13}$
120	$-2b_{12}a_{13} - d_{12}d_{13} + b_{23}e_{13} + d_{23}f_{13}$
021	$f_{13}f_{23} - b_{12}e_{13} - d_{13}f_{12} + 2c_{23}c_{13}$
102	$2c_{13}e_{23} + e_{13}c_{23} - 2a_{13}c_{12} - e_{13}e_{12}$
012	$2c_{13}f_{23} + f_{13}c_{23} - d_{13}c_{12} - e_{13}f_{12}$
111	$e_{13}f_{23} + f_{13}e_{23} - d_{12}e_{13} - d_{13}e_{12} - 2a_{13}f_{12} + 2d_{23}c_{13}$

$D_2 w_{23}$

300,030,003	$-2a_{12}a_{23} + a_{23}e_{23}, -b_{12}d_{23} + b_{23}f_{23}, -e_{23}c_{12} + 2c_{23}^2$
210	$-a_{12}d_{23} - 2d_{12}a_{23} + a_{23}f_{23} + d_{23}e_{23}$
201	$e_{23}^2 - a_{12}e_{23} - 2a_{23}e_{12} + 2a_{23}c_{23}$
120	$-2b_{12}a_{23} - d_{12}d_{23} + b_{23}e_{23} + d_{23}f_{23}$
120	$f_{23}^2 - b_{12}e_{23} - f_{12}d_{23} + 2b_{23}c_{23}$
102	$-2a_{23}c_{12} - e_{23}e_{12} + 3c_{23}e_{23}$
012	$-c_{12}d_{23} - e_{23}f_{12} + 3c_{23}f_{23}$
111	$-d_{12}e_{23} - e_{12}d_{23} - 2a_{23}f_{12} + 2d_{23}c_{23} + 2e_{23}f_{23}$

D_3w_{12}	
300,030,003	$-2a_{12}a_{13} - d_{12}a_{23}, -2b_{12}b_{23} - d_{12}b_{13}, -c_{13}e_{12} - c_{23}f_{12}$
210	$-d_{12}d_{23} - 2a_{12}d_{13} - d_{12}a_{13} - 2b_{12}a_{23}$
201	$-2a_{12}e_{13} - a_{13}e_{12} - a_{23}f_{12} - d_{12}e_{23}$
120	$-2b_{12}d_{23} - d_{12}b_{23} - d_{12}d_{13} - 2a_{12}b_{13}$
021	$-2b_{12}f_{23} - f_{12}b_{23} - b_{13}e_{12} - d_{12}f_{13}$
102	$-e_{13}e_{12} - 2c_{13}a_{12} - c_{23}d_{12} - e_{23}f_{12}$
012	$-f_{12}f_{23} - c_{13}d_{12} - f_{13}e_{12} - 2c_{23}b_{12}$
111	$-d_{12}f_{23} - f_{12}d_{23} - d_{12}e_{13} - d_{13}e_{12} - 2a_{12}f_{13} - 2b_{12}e_{23}$

D_3w_{13}	
300,030,003	$-2a_{13}^2 - a_{23}d_{13}, -b_{13}d_{13} - 2b_{23}b_{13}, -c_{13}e_{13} - f_{13}c_{23}$
210	$-3a_{13}d_{13} - 2a_{23}b_{13} - d_{23}d_{13}$
201	$-3a_{13}e_{13} - a_{23}f_{13} - d_{13}e_{23}$
120	$-d_{13}^2 - 2b_{13}a_{13} - b_{23}d_{13} - 2d_{23}b_{13}$
021	$-b_{13}e_{13} - d_{13}f_{13} - b_{23}f_{13} - 2b_{13}f_{23}$
102	$-e_{13}^2 - 2a_{13}c_{13} - d_{13}c_{23} - f_{13}e_{23}$
012	$-d_{13}c_{13} - e_{13}f_{13} - 2b_{13}c_{23} - f_{13}f_{23}$
111	$-2d_{13}e_{13} - 2a_{13}f_{13} - d_{23}f_{13} - 2b_{13}e_{23} - d_{13}f_{23}$

D_3w_{23}	
300,030,003	$-2a_{13}a_{23} - a_{23}d_{23}, -2b_{23}^2 - d_{23}b_{13}, -c_{13}e_{23} - c_{23}f_{23}$
210	$-d_{23}^2 - a_{13}d_{23} - 2a_{23}d_{13} - 2a_{23}b_{23}$
201	$-a_{13}e_{23} - 2a_{23}e_{13} - a_{23}f_{23} - d_{23}e_{23}$
120	$-2a_{23}b_{13} - d_{23}d_{13} - 3b_{23}d_{23}$
021	$-b_{13}e_{23} - d_{23}f_{13} - 3b_{23}f_{23}$
102	$-2a_{23}c_{13} - e_{13}e_{23} - d_{23}c_{23} - e_{23}f_{23}$
012	$-f_{23}^2 - d_{23}c_{13} - f_{13}e_{23} - 2b_{23}c_{23}$
111	$-d_{13}e_{23} - d_{23}e_{13} - 2a_{23}f_{13} - 2d_{23}f_{23} - 2b_{23}e_{23}$

We have the expression $c_{12}(f_{12} + 2c_{13})$, which is $D_1w_{12}003$.

§2.3 Ω is constant

Our next task is showing that Ω is in fact constant matrix.

Since E is finite dimensional with maximal rank, there exist constants c_i 's such that $x_i + c_i \in E$ for $i = 1, 2, 3$. Then we have the following elements in E : $D_j = [L_0, x_j + c_j]$, $\omega_{ji} = [D_i, D_j]$, and

$$[[L_0, D_j], D_k] = \sum_{i=1}^3 \left(\omega_{ji}\omega_{ki} - \frac{\partial \omega_{ji}}{\partial x_k} D_i \right) - \frac{1}{2} \sum_{i=1}^3 \frac{\partial^2 \omega_{ji}}{\partial x_k \partial x_i} - \frac{1}{2} \frac{\partial^2 \eta}{\partial x_k \partial x_j}$$

As ω_{ji} is linear, so we may infer that

$$\sum_{i=1}^3 \omega_{ji}\omega_{ki} - \frac{1}{2} \frac{\partial^2 \eta}{\partial x_k \partial x_j} \in E$$

More compactly we write

$$\Omega\Omega^T - \frac{1}{2}\text{Hess}(\eta) \quad \text{or} \quad \Omega^2 + \frac{1}{2}\text{Hess}(\eta)$$

which is a matrix with entries in E . Ocone's result says that these entries are at most quadratic polynomials. So η is a polynomial of degree at most 4.

If E possesses a nontrivial quadratic polynomial, the argument is very involved and we'll not discuss it here. We consider the case that E has at most linear polynomials.

Writing $P_i = \{\text{polynomials with degree at most } i\}$.

Let $\Omega = Ax_1 + Bx_2 + Cx_3 \pmod{P_0}$ where $A = (a_{ij})_{3 \times 3}$, $B = (b_{ij})_{3 \times 3}$, $C = (c_{ij})_{3 \times 3}$ are skew-symmetric matrices. We make use of $\Omega^2 + \frac{1}{2}\text{Hess}(\eta) = 0 \pmod{P_1}$ to infer that $A = B = C = (0)_{3 \times 3}$ as follows. Let

$$\begin{aligned} H &= \Omega^2 = H_{11}x_1^2 + H_{22}x_2^2 + H_{33}x_3^2 + H_{12}x_1x_2 + H_{13}x_1x_3 + H_{23}x_2x_3 \\ &= A^2x_1^2 + B^2x_2^2 + C^2x_3^2 + (AB + BA)x_1x_2 + (AC + CA)x_1x_3 + (BC + CB)x_2x_3 \end{aligned}$$

We consider terms in η and relationships derived from $\Omega^2 + \frac{1}{2}\text{Hess}(\eta) = 0 \pmod{P_1}$ in terms of entries in H_{ij} matrices. Coefficient of $x_1^2x_2^2$ in $-\eta = H_{11}[2, 2] = H_{22}[1, 1] = \frac{1}{2}H_{12}[1, 2]$. Similarly, $H_{11}[3, 3] = H_{33}[1, 1] = \frac{1}{2}H_{13}[1, 3]$ ($H_{ij}[p, q]$ means the (p, q) -entry of matrix H_{ij}). We have :

$$(2.2) \quad a_{12}^2 + a_{23}^2 = b_{12}^2 + b_{13}^2 = \frac{1}{2}(a_{13}b_{23} + a_{23}b_{13})$$

$$(2.3) \quad a_{13}^2 + a_{23}^2 = c_{12}^2 + c_{13}^2 = -\frac{1}{2}(a_{12}c_{23} + a_{23}c_{12})$$

Suppose A, B, C are not all zero matrices, without loss of generality we may assume that $A \neq O_{3 \times 3}$:

Consider terms $x_1^2 \sum_{j,k \neq 1} x_j x_k$ in η . There exists an orthogonal transformation R leaving x_1 fixed and changes $\sum_{j,k \neq 1} x_j x_k$ to canonical form $k_2 \tilde{x}_2^2 + k_3 \tilde{x}_3^2$ where $\tilde{x} = Rx$. So without loss of generality we may assume from the beginning that η does not contain the term $x_1^2 x_2 x_3$, which implies that $H_{11}[2, 3] = 0$. Hence $a_{12}a_{13} = 0$.

Suppose $a_{13} = 0$. (2.2) implies

$$\begin{aligned} b_{12}^2 + b_{13}^2 + a_{12}^2 + a_{23}^2 &= a_{23}b_{13} \leq \frac{1}{2}(a_{23}^2 + b_{13}^2) \\ \Rightarrow 2b_{12}^2 + b_{13}^2 + 2a_{12}^2 + a_{23}^2 &\leq 0 \\ \Rightarrow a_{12} &= a_{23} = 0 \end{aligned}$$

So $A = O_{3 \times 3}$. Contradiction. Similarly, $a_{12} = 0$ will lead to a contradiction also. Hence A, B, C are simultaneously zero matrices.

For reference we list out the H_{ij} matrices below :

$$\begin{aligned} H_{11} &= - \begin{pmatrix} a_{12}^2 + a_{13}^2 & a_{13}a_{23} & a_{12}a_{23} \\ a_{13}a_{23} & a_{12}^2 + a_{23}^2 & a_{12}a_{13} \\ a_{12}a_{23} & a_{12}a_{13} & a_{13}^2 + a_{23}^2 \end{pmatrix} \\ H_{22} &= - \begin{pmatrix} b_{12}^2 + b_{13}^2 & b_{13}b_{23} & b_{12}b_{23} \\ b_{13}b_{23} & b_{12}^2 + b_{23}^2 & b_{12}b_{13} \\ b_{12}b_{23} & b_{12}b_{13} & b_{13}^2 + b_{23}^2 \end{pmatrix} \\ H_{33} &= - \begin{pmatrix} c_{12}^2 + c_{13}^2 & c_{13}c_{23} & c_{12}c_{23} \\ c_{13}c_{23} & c_{12}^2 + c_{23}^2 & c_{12}c_{13} \\ c_{12}c_{23} & c_{12}c_{13} & c_{13}^2 + c_{23}^2 \end{pmatrix} \\ H_{12} &= - \begin{pmatrix} 2a_{12}b_{12} + 2a_{13}b_{13} & a_{13}b_{23} + a_{23}b_{13} & -a_{12}b_{23} - a_{23}b_{12} \\ a_{13}b_{23} + a_{23}b_{13} & 2a_{12}b_{12} + 2a_{23}b_{23} & a_{12}b_{13} + a_{13}b_{12} \\ -a_{12}b_{23} - a_{23}b_{12} & a_{12}b_{13} + a_{13}b_{12} & 2a_{13}b_{13} + 2a_{23}b_{23} \end{pmatrix} \\ H_{13} &= - \begin{pmatrix} 2a_{12}c_{12} + 2a_{13}c_{13} & a_{13}c_{23} + a_{23}c_{13} & -a_{12}c_{23} - a_{23}c_{12} \\ a_{13}c_{23} + a_{23}c_{13} & 2a_{12}c_{12} + 2a_{23}c_{23} & a_{12}c_{13} + a_{13}c_{12} \\ -a_{12}c_{23} - a_{23}c_{12} & a_{12}c_{13} + a_{13}c_{12} & 2a_{13}c_{13} + 2a_{23}c_{23} \end{pmatrix} \\ H_{23} &= - \begin{pmatrix} 2b_{12}c_{12} + 2b_{13}c_{13} & b_{13}c_{23} + b_{23}c_{13} & -b_{12}c_{23} - b_{23}c_{12} \\ b_{13}c_{23} + b_{23}c_{13} & 2b_{12}c_{12} + 2b_{23}c_{23} & b_{12}c_{13} + b_{13}c_{12} \\ -b_{12}c_{23} - b_{23}c_{12} & b_{12}c_{13} + b_{13}c_{12} & 2b_{13}c_{13} + 2b_{23}c_{23} \end{pmatrix} \end{aligned}$$

References

- [0] V. BENES, Exact finite dimensional filters for certain diffusions with nonlinear drift, *Stochastics*, 5 (1981), pp. 65-92.
- [1] R.W. BROCKETT AND J.M.C. CLACK, The geometry of the conditional density functions, in *Analysis and Optimization of Stochastic Systems*, O.L.R. Jacobs et al., eds., Academic Press, New York, 1980, pp. 299-309.
- [2] R.W. BROCKETT, Nonlinear systems and nonlinear estimation theory, in *The Mathematics of Filtering and Identification and Applications*, M. Hazewinkel and J.S. Willems, eds., Reidel, Dordrecht, 1981.
- [3] R.W. BROCKETT, Nonlinear Control Theory and Differential Geometry, *Proceedings of the International Congress of Mathematics*, (1983), pp. 1357-1368.
- [4] M. CHALEYAT-MAUREL AND D. MICHEL, Des resultats de non-existence de filtre de dimension finie, *Stochastics*, 13 (1984), pp. 83-102.
- [5] W.L. CHIOU AND S.S.-T. YAU, Finite dimensional filters with nonlinear drift II: Brockett's problem on Classification of finite dimensional estimation algebras (submitted).
- [6] P.C. COLLINGWOOD, Some remarks on estimation algebras, *Systems Control Lett.*, 7 (1986), pp. 217-224.
- [7] M.H.A. DAVIS, On a multiplicative functional transformation arising in nonlinear filtering theory, *Z. Wahrsch. Verw. Gebiete*, 54 (1980), pp. 125-139.
- [8] M.H.A. DAVIS AND S.I. MARCUS, An introduction to nonlinear filtering, in *The Mathematics of Filtering and Identification and Applications*, M. Hazewinkel and J.S. Willems, eds., Reidel, Dordrecht, 1981.
- [9] R.T. DONG, L.F. TAM, W.S. WONG and S. S.-T. YAU, Structure and Classification Theorems of Finite Dimensional Exact Estimation Algebras, *SIAM J. Control and Optimization*, Vol.29, No.4, pp. 866-877, July 1991
- [10] M. FUJISAKI, G. KALLIANPUR AND H. KUNITA, Stochastic Differential Equations for the Nonlinear Filtering Problem, *Osaka J. of Math.*, Vol.1 (1972), pp. 19-40.
- [11] S.K. MITTER, On the analogy between mathematical problems of nonlinear filtering and quantum physics, *Ricerche di Automatica* 10 (2) (1979), pp.163-216.
- [12] D.L. OCONE, Finite dimensional estimation algebras in nonlinear filtering, in *The Mathematics of Filtering and Identification and Applications*, M. Hazewinkel and J.S. Willems, eds., Reidel, Dordrecht, 1981..
- [13] S. STEINBERG, Applications of the Lie algebraic formulas of Baker, Campbell, Hausdorff and Zassenhaus to the calculation of explicit solutions of partial differential equations, *J. Differential Equations*, 26 (1979), pp. 404-434.
- [14] L.F. TAM, W.S. WONG and S. S.-T. YAU, On a necessary and sufficient condition for finite dimensionality of estimation algebras, *SIAM J. Control and Optimization*, Vol. 28, No. 1 (1990), pp. 173-185.
- [15] J. WEI and E. NORMAN, On global representation of the solutions of linear differential equations as a product of exponentials, *Proc. Amer. Math. Soc.*, 15 (1964), pp. 327-334.
- [16] D.V. WIDDER, *The Heat Equation*, Mathematics 67, Academy Press, 1975
- [17] W.S. WONG, New classes of finite dimensional nonlinear filters, *Systems Control Lett.*, 3 (1983), pp. 155-164.
- [18] W.S. WONG, On a new class of finite dimensional estimation algebras, *Systems Control Lett.*, 9 (1987), pp. 79-83.
- [19] W.S. WONG, Theorems on the structure of finite dimensional estimation algebras, *Systems Control Lett.*, 9 (1987), pp. 117-124.
- [20] S. S.-T. YAU, Recent results on nonlinear filtering, New class of finite dimensional filters, *Proceedings of the 29th IEEE Conference on Decision and Control*, Honolulu, Hawaii, Dec.5-7, (1990), pp. 231-233.
- [21] S. S.-T. YAU, Finite dimensional filters with nonlinear drift I : A class of filters including both Kalman-Bucy filters and Benes filters (to appear) *J. of Math. Systems, Estimation and Control*
- [22] S. S.-T. YAU AND W.L. CHIOU, Recent results on classification of finite dimensional estimation algebras : Dimension of State Space ≤ 2 , *Proceedings of the 30th IEEE Conference on Decision and Control*, Brighton, England, Dec 11-13, (1991), pp. 2758-2760.

Hybrid Optimal Control of Turret-Gun System†

J. L. Zhang, L. S. Shieh

Electrical Engineering Department

University of Houston

4800 Calhoun

Houston, TX 77204-4793

N. P. Coleman

Army Research, Development, and Engineering Center

SMCAR-FSF-RC, Bldg. 95 North

Dover, NJ 07806-5000

Abstract

This paper presents a hybrid technique for optimal discrete-time control of a continuous-time Turret-gun system. The optimal regional-pole placement technique is utilized to design a continuous-time linear state-feedback control law. This law is then converted to an equivalent discrete-time control law for digital implementation aided by the recently developed digital redesign technique and ideal state reconstructor. A preload compensation is also added to the digital controller for reducing the effect of the turret motor friction. A digital simulation of the designed nonlinear Turret-gun system is presented.

1. Introduction

The Turret-gun system contains hard nonlinearities such as Coulomb friction, backlash, and saturation, and it is subject to external disturbances such as the firing disturbances and base motion, as well as, parameter variations caused by thermal effects on gun

† supported by the U.S. Army Research Office under contract DAAL-03-91-G0106 and the NASA-Johnson Space Center under grant NAG-9-380

barrel and torsional stiffness. The primary objective of the Turret-gun servo control system is to rapidly stabilize and accurately point the gun to a target position in the presence of the above nonlinearities, disturbances, and model uncertainties. This requires that the designed servo control system has the properties of robust, rapid and accurate tracking, and disturbance and noise rejection. To develop, demonstrate, and validate the advanced algorithms for Turret-gun control system, the advanced weapon tracking testbed (ATB-1000) was designed at the Army Research Development Engineering Center to provide a realistic simulation of Turret-gun system in laboratory environment [1]. The detailed linear and nonlinear models, developed by Integrated System Inc. (ISI), are used as the basis for this study.

The primary objective of this paper is to develop a digitally implementable optimal controller for the continuous-time nonlinear Turret-gun system so that the responses of the system converge at an appropriate speed and any vibrating modes are well damped. The hybrid controller design methodology developed in this paper can be briefly described as follows.

A cascaded internal model (which contains an integrator) and a feedback integrator are inserted into the continuous-time linearized model to reduce nonlinear effects and firing disturbances. Based on the above integrated linear model, a linear quadratic regulator approach [2] is utilized to design a robust optimal state-feedback control law. This law optimally places the closed-loop poles of the above integrated linear model within the common region of an open sector and to the left hand side of a line parallel to the imaginary axis in the complex s -plane. Thus, the designed system responses converge at an appropriate speed and any vibrating modes are well damped. Moreover, for digital control of the Turret-gun system, recently developed digital redesign techniques [3] are applied for digital redesign of the inserted continuous-time internal model and the developed continuous-time optimal state-feedback control law. In order to implement these digitally redesigned controllers without constructing a digital observer, a new ideal state reconstructor is developed

for the estimation of the discretized states under noise disturbances and nonlinear effects.

2. Design formulation

A linear model obtained from the given nonlinear model of the ATB-1000 test fixture is described by following state space equation,

$$\begin{aligned}\dot{x}_o &= A_o x_o + B_o u_o \\ y_o &= C_o x_o\end{aligned}\tag{1}$$

where $x_o \in \mathcal{R}^{10 \times 1}$, $u_o \in \mathcal{R}$, $y_o \in \mathcal{R}^{6 \times 1}$, and A_o , B_o , C_o are matrices with appropriate dimensions. The effect of nonlinearities is simply ignored in this model, but these will be considered in the formulation of the controller structure. The six outputs of the system are turret motor yaw, turret motor rate, inertial wheel yaw, strain gauge 1, strain gauge 2, and tip acceleration, respectively. Although there are other measurable outputs on the ATB-1000 test fixture, they are not practically accessible for control of an actual Turret-gun system.

In order to reduce the effect of nonlinearities and to increase the robustness of the designed control system, an internal model which contains an integrator is inserted before the system. The steady-state error of the barrel tip position caused by nonlinearities is eliminated by another integrator appended to third output of y_o , i.e., the inertia wheel yaw. Note that the inertia wheel yaw is used here instead of the barrel tip position due to the non-availability of tip sensor. With the addition of these two integrators, the overall system is augmented to an 12th order system as shown in Fig. 1. The augmented state space equation is given as

$$\begin{aligned}\dot{x} &= Ax + Bu \\ y &= Cx\end{aligned}\tag{2}$$

where

$$x = \begin{bmatrix} x_{i1} \\ x_o \\ x_{i2} \end{bmatrix} \in \mathcal{R}^{12 \times 1}, u \in \mathcal{R}, y \in \mathcal{R}$$

$x_{i1} \triangleq$ output of the first integrator

$x_{i2} \triangleq$ output of the second integrator

$$A = \begin{bmatrix} 0 & 0_{1 \times 10} & 0 \\ B_o & A_o & 0_{10 \times 1} \\ 0 & -C_{o3} & 0 \end{bmatrix} \in \mathcal{R}^{12 \times 12}$$

$$B = \begin{bmatrix} 1 \\ 0_{10 \times 1} \\ 0 \end{bmatrix} \in \mathcal{R}^{12 \times 1}, C = [0 \quad 0_{1 \times 10} \quad 1] \in \mathcal{R}^{1 \times 12}$$

The r in Fig. 1 is the reference input or the command input for the inertia wheel yaw (which is identical to the tip position in steady state). Since r is considered as an external disturbance in the design of the optimal state feedback control law, it does not appear in (2)

The eigenvalues of the system in (2) are $\{0.0, 0.0, -4.50, -2.11 \pm j27.57, -5.76 \pm j59.61, -2.52 \pm j138.43, -5.77 \pm j384.88\}$. Thus, the eigenvalues have a large spread. Although the optimal regional-pole placement method can be directly used to optimally place the closed-loop poles in the previously mentioned region, a rather large feedback gain would be obtained as a result of the large eigenvalue spread. This is undesirable due to the presence of nonlinearities and unmodelled dynamics. Since the state feedback gain is to be connected to the observed or reconstructed states, any deviation of the observed or reconstructed state from the actual state will significantly affect the control effort which would saturate the actuator and, even worse, destabilize the closed-loop system. Therefore, it is advantageous to decompose this large-scale and stiff system into a completely decoupled multi-time scale structure, so that each subsystem has its own distinct characteristics and can be designed accordingly.

The multi-stage design algorithm developed by Tsai, *et al.* [2] is modified and utilized here to design the system in (2). The design procedures can be described as follows:

Step 1. The eigenvalues of the augmented system in (2) are divided into three groups that are located in three circular rings, $[0, r_1], [r_1, r_2], [r_2, r_3]$, as shown in Fig. 2, where $r_1 = 15, r_2 = 100, r_3 = 400$. The first circular ring contains the eigenvalues $\{0, 0, 0, -4.5\}$; the second circular ring $\{-2.11 \pm j27.57, -5.76 \pm j59.61\}$; and the third circular ring $\{-2.52 \pm j138.43, -5.77 \pm j384.88\}$. A block modal matrix is constructed by using the matrix sign algorithm [3] to decompose the original system into three decoupled subsystems corresponding to the three groups of eigenvalues. The modal matrix is given as

$$M_s = [S_3 \ S_2 \ S_1] \quad (3)$$

where

$$S_i \triangleq \text{ind}[\text{sign}_{(r_{i-1}, r_i)}^+(h(A))] \in \mathcal{R}^{12 \times n_i}, \ 1 \leq i \leq 3$$

$$\text{and } n_1 = 4, n_2 = 4, n_3 = 4$$

$$\text{sign}_{(r_{i-1}, r_i)}^+(h(A)) \triangleq \frac{1}{2}[\text{sign}_{(r_{i-1})}(h(A)) - \text{sign}_{(r_i)}(h(A))]$$

$$h(A) = (A - r_i I_n)(A + r_i I_n)^{-1}$$

In the above definitions, $\text{ind}(\cdot)$ represents the collection of the linearly independent column vectors of (\cdot) and $r_0 = 0, \text{sign}_{(0)}(h(A)) = I_n$. Then

$$\begin{aligned} A_d &= M_s^{-1} A M_s = \text{block diag}(A_{d3}, A_{d2}, A_{d1}) \\ B_d &= M_s^{-1} B = \begin{bmatrix} B_{d3} \\ B_{d2} \\ B_{d1} \end{bmatrix} \end{aligned} \quad (4)$$

where $A_{di} \in \mathcal{R}^{4 \times 4}, B_{di} \in \mathcal{R}^{4 \times 1}$.

Step 2: Set $i = 1, \bar{A} = A_d = \text{block diag}[\bar{A}_3, \bar{A}_2, \bar{A}_1], \bar{B} = B_d = \text{block diag}[B_3^T, B_2^T, B_1^T]^T, M_1 = I_n$, and the feedback gain $\bar{K}_c = 0_{m \times n}$.

Step 3: The subsystem considered for design at this stage is (\bar{A}_i, \bar{B}_i) . Design this subsystem by using the optimal pole placement method [2]. Let the immediate optimal feedback gain be \bar{K}_i and the corresponding continuous-time closed-loop system be $(\bar{A}_{ci}, \bar{B}_i)$.

Step 4: Update

$$\bar{K}_c := \bar{K}_c + [0_{m \times (n-n_i)}, \bar{K}_i] M_1^{-1} \quad (5a)$$

$$\bar{A} := \bar{A} - \bar{B}[0_{m \times (n-n_i)}, \bar{K}_i] = \begin{bmatrix} \bar{A}_i & W_i \\ 0_{n_i \times (n-n_i)} & \bar{A}_{c_i} \end{bmatrix} \quad (5b)$$

where $\bar{A}_i = \text{block diag} [\hat{A}_{c_i}, \hat{A}_i]$, $W_i = -[\bar{B}_i^T, \hat{B}_i^T]^T \bar{K}_i$. The n_i is the order of the subsystem that is being designed at this stage. The dimensions of the matrices \bar{A}_i and W_i are $(n - n_i) \times (n - n_i)$, $(n - n_i) \times n_i$, respectively.

Step 5: Block-diagonalize the partially designed system \bar{A} and move the last block of \bar{A} in (5b) (viz., \bar{A}_{c_i}) to the first block, via a transformation matrix M_2 which is given as

$$M_2 = \begin{bmatrix} L_i & I_{n-n_i} \\ I_{n_i} & 0_{n_i \times (n-n_i)} \end{bmatrix}, \quad M_2^{-1} = \begin{bmatrix} 0_{n_i \times (n-n_i)} & I_{n_i} \\ I_{n-n_i} & -L_i \end{bmatrix} \quad (6a)$$

The matrix $L_i \in \mathcal{R}^{(n-n_i) \times n_i}$ can be solved from the following Lyapunov equation,

$$\bar{A}_i L_i - L_i \bar{A}_{c_i} + W_i = 0_{(n-n_i) \times n_i} \quad (6b)$$

The transformed system is

$$\bar{A} := M_2^{-1} \bar{A} M_2 = \begin{bmatrix} \bar{A}_{c_i} & 0_{n_i \times (n-n_i)} \\ 0_{(n-n_i) \times n_i} & \bar{A}_i \end{bmatrix} \quad (6c)$$

$$\bar{B} := M_2^{-1} \bar{B} = [\bar{B}_i^T, (\bar{B}_i - L_i \bar{B}_i)^T]^T \quad (6d)$$

where $\bar{B}_i = [B_c^T, \hat{B}_i^T]^T$. Accumulate the transformations in $M_1 := M_1 M_2$.

Step 6: Set $i := i + 1$. If $i > k$ (k is the number of time-scales. In this specific case, it equals 3), then go to Step 7; else, go to Step 3.

Step 7: The state-feedback gain for the original system is obtained as $K_c = \bar{K}_c M_1^{-1}$.

The final result is given as follows: with the designing parameters $h_1 = 15, h_2 = 20, h_3 = 15$, the state-feedback gain is computed as

$$K_c = [391.0 \quad 23.0 \quad 144.0 \quad 23.0 \quad 250.0 \quad 11.0 \quad -43.0 \quad 23.0 \quad 241.0 \quad 19.0 \quad 328.3 \quad -41071.0] \quad (7)$$

The closed-loop eigenvalues are given as $\{-25.5, -30.0, -30.0, -30.0, -37.89 \pm j27.57, -66.40 \pm j66.40, -27.48 \pm j138.43, -24.23 \pm j384.88\}$.

3. Digital implementation

3.1. Digital redesign

In order to implement the designed continuous-time state-feedback control law in the sampled-data environment of the ATB-1000 test fixture, the digital redesign technique [3] is used to convert the continuous-time control gain K_c in (7) to an equivalent discrete-time control gain K_d . The state feedback gain K_c can be converted [3] by

$$K_d = \frac{1}{2}(I_m + \frac{1}{2}K_c H)^{-1} K_c (I_n + G) \quad (8)$$

where K_c is given in (7), $G = e^{AT}$, $H = \int_0^T e^{A\tau} d\tau$, and $T = 0.0025$ sec. The K_d is computed as

$$K_d = [K_{d1}, K_{do}, K_{d2}] \quad (9)$$

where

$$K_{d1} = 292.89$$

$$K_{do} = [5.7936, 83.24, 21.30, 164.88, 11.77, -10.91, 12.20, 208.01, -21.46, 1750.4]$$

$$K_{d2} = -24429.0$$

The digital control law is then given as

$$u_d(kT) = -K_d x_d(kT) = -K_{d1} x_{di1}(kT) - K_{do} x_{do}(kT) - K_{d2} x_{di2}(kT) \quad (10)$$

Furthermore, the inserted integrators can be discretized by using the Tustin approximation:

$$\frac{1}{s} \rightarrow \frac{2}{T} \frac{(1+z^{-1})}{(1-z^{-1})}.$$

3.2. Digital state reconstructor using recursive weighted least-squares (RWLS) algorithm

$x_{di1}(kT)$ and $x_{di2}(kT)$ in (10) are outputs of the inserted integrators and are directly accessible. However, the state x_{do} is not directly measurable. Thus, a state reconstructor is developed in this paper to reconstruct the actual state $x_{do}(kT)$ from the measurable outputs and inputs of the system in (1).

Let the discretized state space equation for system in (1) be

$$\begin{aligned}x_{dok+1} &= Gx_{dok} + Hu_{dk} \\ y_{dok} &= Cx_{dok}\end{aligned}\tag{11}$$

Then, the following RWLS algorithm can be used to estimate the digital state in (11):

$$\hat{x}_{k+1} = G\hat{x}_k + Hu_k + K_k(y_k - C\hat{x}_k)\tag{12a}$$

$$K_k = GP_kC^T(\lambda I_p + CP_kC^T)^{-1}\tag{12b}$$

$$P_{k+1} = \frac{1}{\lambda}[G - K_kC]P_kG^T\tag{12c}$$

where \hat{x}_k is the estimation of the digital state x_{do} at $t = kT$, λ is a forgetting factor.

The algorithm presented here bears many resemblances to the Kalman filtering algorithm. The RWLS algorithm is simpler in terms of the required knowledge of noise characteristics, and thus provides a practical alternative to Kalman filtering.

3.3. Friction compensation

It is observed that the Coulomb friction presented in the turret motor has a significant undesirable effect on the overall system performance. The presence of the Coulomb friction makes the system response rather sluggish, and the system response takes much longer time to settle down. Simply increasing the feedback gain does not solve the problem due to the limitation of the actuator. In this paper, an inverse nonlinear technique is utilized to deal with this problem. Specifically, a nonlinear compensation term is added to the input of the turret motor. This compensation term takes the form of $u_{comp}(kT) = f_c \text{sign}(y_{do2}(kT))$, where f_c is the turret motor friction and $y_{do2}(kT)$ is the sampled value of the turret motor velocity. The effect of this nonlinear compensation is to convert the low frequency disturbance caused by the turret motor friction to a high frequency impulse disturbance. The converted high frequency impulse disturbance has less effect on the system due to the lowpass nature of the designed closed-loop system. The control law in (10) now becomes

$$u_d(kT) = -K_{d1}x_{di1}(kT) - K_{do}\hat{x}_{do}(kT) - K_{d2}x_{di2}(kT) + u_{comp}(kT)\tag{13}$$

4. Simulation results

The complete digitally implemented control system of the ATB-1000 test fixture is shown in Fig. 3. Nonlinear simulations are carried out using the MATRIX_x simulation tools and the results are shown in Figs. 4-5. Nonlinear simulation of the closed-loop system using the control law without the friction compensation is also carried out and the results are shown in the same figures for comparison.

5. Conclusions

A hybrid technique for optimal discrete-time control of a continuous-time Turret-gun system has been proposed in this paper. A continuous-time linear state-feedback control law has been designed by using the optimal regional-pole placement technique. This law is then converted to an equivalent discrete-time control law for digital implementation aided by the digital redesign technique. A recursive weighted least-squares algorithm has been utilized to estimate the digital state of the Turret-gun system. A preload compensation is also added to the digital controller for reducing the effect of the turret motor friction. Finally, all these components have been integrated together to yield a practically implementable controller for the ATB-1000 test fixture. The result of the nonlinear digital simulation of the designed nonlinear Turret-gun system has demonstrated the effectiveness of the proposed techniques.

References

- [1] M. Mattice, N. Coleman, S. Banks, J. C. Juang, and C. F. Lin, "Robust weapon control systems design", *Proc. of American Control Conference*, pp.429-433, 1992
- [2] L. S. Shieh, H. M. Dib, and S. Ganesan, "Continuous-time quadratic regulators and pseudo-continuous-time quadratic regulators with pole placement in a specific region," *IEE Proc. Pt. D*, 134(5):338-346, 1987.
- [3] J. S. H. Tsai, L. S. Shieh, J. L. Zhang, and P. C. Coleman, "Digital redesign of pseudo-continuous-time suboptimal regulators for large-scale discrete systems", *Control-Theory*

and Advanced Technology, 5(1):37-65, 1989.

- [4] L. S. Shieh, X. M. Zhao, and J. W. Sunkel, "Hybrid state-space self-tuning control using dual-rate sampling," *IEE Proc. Pt. D*, 138(1):50-58, 1991.

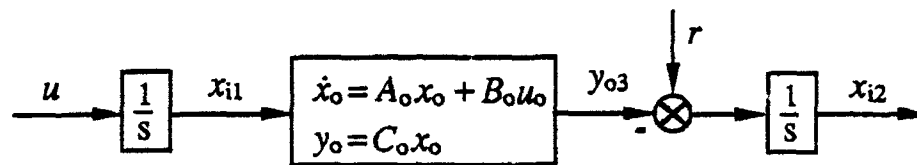


Fig. 1 Augmented system

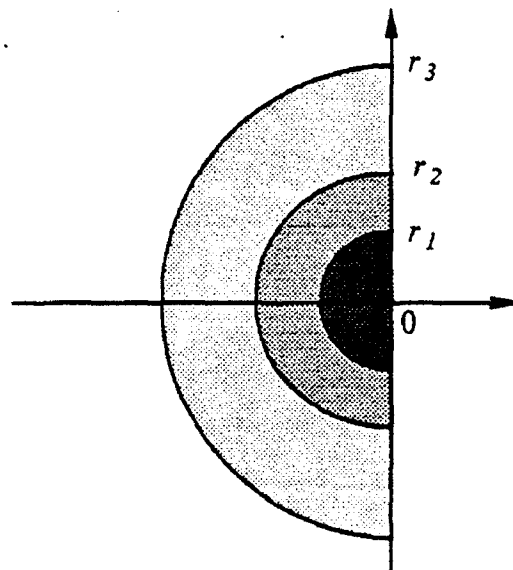


Fig. 2 The three circular rings of interest

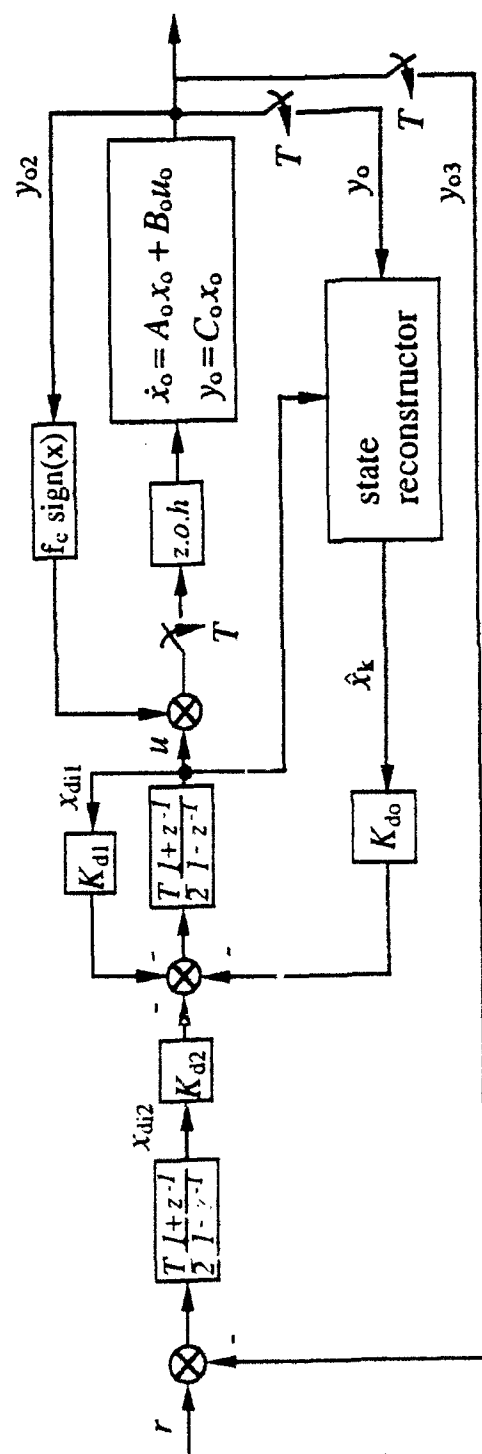


Fig. 3 Designed digitally controlled system

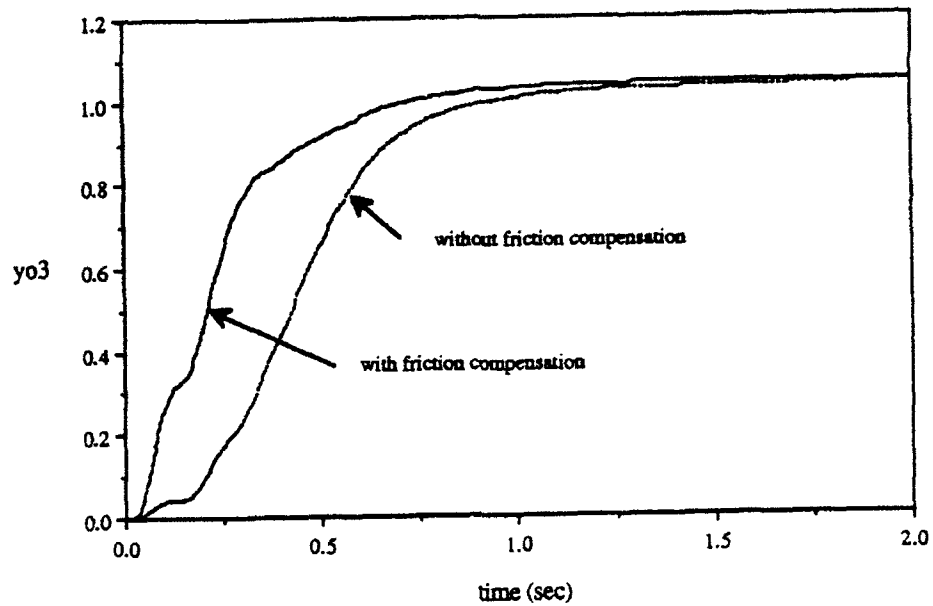


Fig. 4 Step responses of inertial wheel yaw

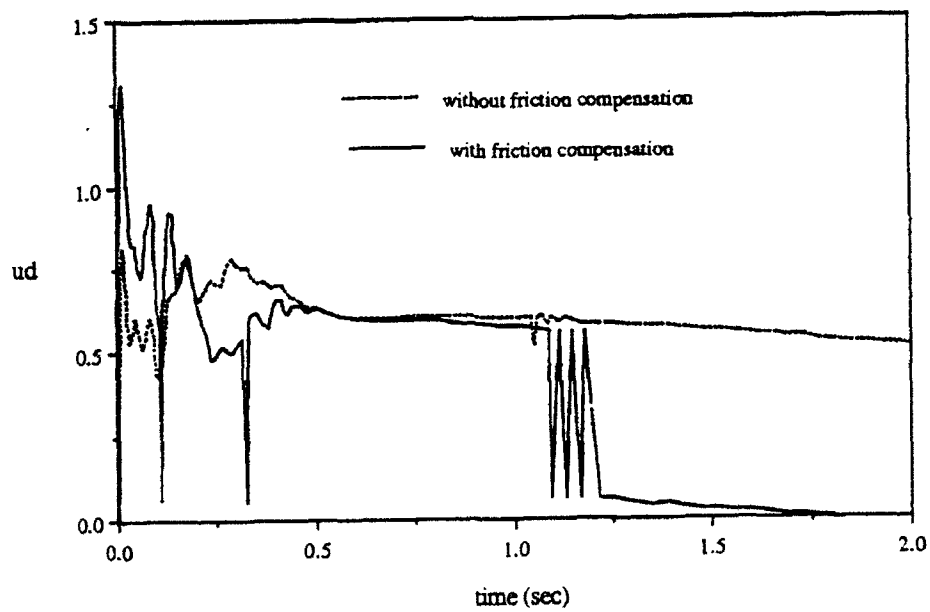


Fig. 5 Control signals applied to turret motor

WAVELET ANALYSIS AND ITS APPLICATIONS*

Charles K. Chui
Center for Approximation Theory
Department of Mathematics
Texas A&M University
College Station, TX 77843-3368

ABSTRACT.

The objective of this writing is to give a brief introduction of the subject of wavelet analysis, and to compare it with the classical subject of Fourier analysis. Spline examples are considered, and the integral wavelet transform is considered as a bandpass filter with variable bandwidth. A comparison of some of the existing wavelets is also given. Finally, a list of applications of wavelets is included. During the hour-talk, demonstrations of these applications were shown by taking advantage of the multimedia facilities.

INTRODUCTION.

While Fourier analysis is a well-established subject within classical analysis and applied mathematics, the subject of wavelet analysis was born only during the last decade. The objective of this writing is to give a comparison of these two subjects. A brief account of the existing wavelets and a list of applications are included in this writing.

1. FOURIER SERIES.

Every 2π -periodic function can be represented by its Fourier series

$$f(x) \sim \sum_{n=-\infty}^{\infty} c_n \psi(nx),$$

* Research partially supported by ARO Contract DAAL 03-90-G-0091.

where a single function

$$\psi(x) := e^{ix} = \cos x + i \sin x$$

is used in the series representation. Not only is this representation very useful, the terms in the series are also very meaningful: namely, each term $c_n \psi(nx)$ indicates the n^{th} "mode" of $f(x)$, when $f(x)$ considered is as a "wave". In this regard, only one "basic" function $\psi(x) = e^{ix}$, is needed to represent all 2π -periodic functions $f(x)$. This function may be considered as the analyzing wave.

2. SERIES REPRESENTATIONS FOR NONPERIODIC FUNCTIONS

For convenience, we only discuss the class $L^2(-\infty, \infty)$ of functions $f(x)$ with "finite energy", namely:

$$\|f\|_2 := \left(\int_{-\infty}^{\infty} |f(x)|^2 dx \right)^{1/2} < \infty.$$

In order to imitate the Fourier series, we again look for a single "basic" function (also denoted by $\psi(x)$) to represent all $f(x) \in L^2(-\infty, \infty)$ in the form of an infinite series as before.

Assuming that $\psi(x)$ should be "Lipschitz continuous", say, then for $\psi(x) \in L^2(-\infty, \infty)$, we have

$$\psi(x) \rightarrow 0, \text{ as } x \rightarrow \pm\infty.$$

So, just "dilation" (i.e. $\psi(nx)$) alone cannot do the job, and we need "translation" also. In addition, in order to group ranges of frequencies (i.e. "frequency bands" or "octaves"), we consider:

"dilation by powers of 2"

instead of dilation by integers. Also, for simplicity, we only consider translations by "integers". That is, we will consider

$$\psi_{j,k}(x) := 2^{j/2} \psi(2^j x - k), \quad j, k \in \mathbf{Z}.$$

Where, as usual, \mathbf{Z} denotes the set of all integers. The normalization constant $2^{j/2}$ is used here so that

$$\|\psi_{j,k}\|_2 = \|\psi\|_2, \quad \text{all } j, k \in \mathbf{Z}.$$

Hence, we are interested in series representations of the type:

$$f(x) \sim \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} c_{j,k} \psi_{j,k}(x)$$

for all $f(x) \in L^2(-\infty, \infty)$. Let the above representation of $f(x)$ be separated into components:

$$\begin{cases} f(x) = \sum_{j=-\infty}^{\infty} g_j(x), & \text{with} \\ g_j(x) = \sum_{k=-\infty}^{\infty} c_{j,k} \psi_{j,k}(x). \end{cases}$$

We see that $g_j(x)$ represents the component of $f(x)$ in the " 2^j th frequency band" (or j th octave). Suppose that the separation

$$f(x) = \sum_{j=-\infty}^{\infty} g_j(x) = \cdots + g_{-1}(x) + g_0(x) + \cdots$$

is "unique" (that is, we have an infinite **direct-sum** decomposition of any $f(x) \in L^2(-\infty, \infty)$). Then by setting

$$W_j := \{g_j: f \in L^2(-\infty, \infty)\},$$

we have a direct-sum decomposition of $L^2(-\infty, \infty)$, namely:

$$L^2(-\infty, \infty) = \bigoplus_{j=-\infty}^{\infty} W_j.$$

Next, let us consider the partial direct sums:

$$V_n := \bigoplus_{j=-\infty}^{n-1} W_j.$$

Then we arrive at a nested sequence of approximating subspaces of $L^2(-\infty, \infty)$, namely:

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots \subset L^2(-\infty, \infty), \quad \text{and}$$

$$\text{clos}_{L^2} \bigcup_{j \in \mathbb{Z}} V_j = L^2(-\infty, \infty)$$

In particular, for each integer n , we also have

$$V_{n+1} = V_n \oplus W_n.$$

That is, the subspaces W_n are complementary subspaces of the nested sequence $\{V_j\}$.

3. EXAMPLE (CARDINAL CUBIC SPLINES).

Let V_0 denote the collection of functions $f(x) \in L^2(-\infty, \infty)$ satisfying:

- (a) $f(x) \in C^2(-\infty, \infty)$; that is, $f(x)$, $f'(x)$, and $f''(x)$ are continuous for all x , and
- (b) the restrictions of $f(x)$ on each of the intervals $[k, k+1]$, $k = \dots, 0, 1, \dots$, are cubic polynomials.

Then each $f(x) \in V_0$ is called a *cardinal cubic spline with knot sequence \mathbf{Z}* .

For each $j \in \mathbf{Z}$, consider

$$V_j = \{f(2^j x) : f(x) \in V_0\}.$$

Then every $f(x) \in V_j$ is a cardinal cubic spline with knots $2^{-j}\mathbf{Z}$. Since removing a knot is equivalent to imposing a third continuous differentiability condition at the knot, we have a nested sequence of subspaces of $L^2(-\infty, \infty)$, namely:

$$\dots \subset V_{-1} \subset V_0 \subset V_1 \subset \dots \subset L^2(-\infty, \infty).$$

From this nested sequence of subspaces V_n , there are many choices of the complementary subspaces W_n .

Let us choose those complementary subspaces as dictated by a given **projection operator**.

(i) **L^2 -projection.**

Let $P_j: L^2(-\infty, \infty) \rightarrow V_j$ be defined by

$$\|f - P_j f\|_2 = \min_{h \in V_j} \|f - h\|_2$$

for all $f \in L^2(-\infty, \infty)$. So, by setting

$$W_j = \{f - P_j f : f \in V_{j+1}\},$$

we have

$$V_{j+1} = V_j \oplus W_j.$$

In fact, this direct sum is an orthogonal sum: $V_j \perp W_j$. Hence,

$$L^2(-\infty, \infty) = \bigoplus_{j=-\infty}^{\infty} W_j.$$

Now, considering the "spline approximation" problem:

$$f \sim f_j := P_j f \in V_j,$$

we then have

$$g_j := f_{j+1} - f_j \sim f - f = 0$$

for all sufficiently large j , provided that f is sufficiently "smooth". But $g_j \in W_j$. So, the j^{th} octave of f is ≈ 0 in very high frequency ranges.

How about if f is smooth on some intervals, but not so smooth elsewhere?

Since the cardinal splines are "locally generated" (by translates of a B -spline), we expect to have

$$\begin{cases} g_j(x) \sim 0 \text{ where } f(x) \text{ is "very smooth";} \\ g_j(x) \text{ reveals the details of } f(x), \text{ elsewhere.} \end{cases}$$

(ii) **Interpolation at the knots.**

Let $I_j: C \cap L^2(-\infty, \infty) \rightarrow V_j$ be defined by

$$(I_j f) \left(\frac{k}{2^j} \right) = f \left(\frac{k}{2^j} \right), \quad k \in \mathbb{Z},$$

for all $f \in C \cap L^2(-\infty, \infty)$.

Again by setting

$$W_j = \{f - I_j f: f \in V_{j+1}\},$$

we have

$$V_{j+1} = V_j \oplus W_j,$$

and hence,

$$C \cap L^2(-\infty, \infty) = \bigoplus_{j=-\infty}^{\infty} W_j.$$

But this direct-sum decomposition is not an orthogonal decomposition.

It is well known that every cubic spline $f(x) \in V_0$ has the representation

$$f(x) = \sum_{k=-\infty}^{\infty} c_k N_4(x - k)$$

where $\{c_k\}$ is square-summable

$$\left(\text{i.e. } \sum_{k=-\infty}^{\infty} |c_k|^2 < \infty \right)$$

and $N_4(x)$ is the cubic B -spline. Consider the orthogonal complementary spaces W_j ; that is,

$$V_1 = V_0 \oplus^\perp W_0$$

defined by using L^2 -projection. Then for every $g \in W_0$, we also have the representation

$$g(x) = \sum_{k=-\infty}^{\infty} d_k \psi_4(x - k)$$

where $\{d_k\}$ is square-summable, and $\psi_4(x)$ is the cubic B -wavelet (see [1] and Figure 2).

While $N_4 \in V_0$ has minimum support in V_0 , $\psi_4 \in W_0$ also has minimum support in W_0 . From W_0 , we can go to any W_j , $j \in \mathbb{Z}$ by defining, as before,

$$\psi_{4;j,k}(x) := 2^{j/2} \psi_4(2^j x - k).$$

Then every $g_j \in W_j$ has a (unique) representation

$$g_j(x) = \sum_{k=-\infty}^{\infty} d_k^j \psi_{4;j,k}(x).$$

Since every $f(x) \in L^2(-\infty, \infty)$ can be separated as a direct sum of $g_j(x) \in W_j$, $j \in \mathbb{Z}$, we have the following **cubic B -wavelet series representation** of $f(x)$:

$$f(x) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} d_k^j \psi_{4;j,k}(x).$$

4. WAVELET SERIES.

Let $\psi \in L^2(-\infty, \infty)$ and set

$$\psi_{j,k}(x) := 2^{j/2} \psi(2^j x - k), \quad j, k \in \mathbb{Z}.$$

Assume that we have a ψ such that every $f(x) \in L^2(-\infty, \infty)$ has the (unique) series expansion:

$$f(x) \sim \sum_{j,k \in \mathbb{Z}} d_k^j \psi_{j,k}(x).$$

Remarks.

- (i) **Riesz Basis.** For implementation, we would rather work with the coefficient sequence $\{d_k^j\}$, $j, k \in \mathbb{Z}$, instead of $f(x)$. This requires "stability"; that is, the existence of positive constants A and B (independent of $f(x)$) such that

$$A \sum_{j,k} |c_k^j|^2 \leq \|f\|_2^2 \leq B \sum_{j,k} |c_k^j|^2.$$

If this stability condition is satisfied, we call $\{\psi_{j,k}\}$ a Riesz basis of $L^2(-\infty, \infty)$.

(ii) For a Fourier series, the Fourier coefficients can be expressed in terms of the 2π -periodic function by using the orthogonality property of $\{e^{inx}\}$, namely:

$$\begin{cases} f(x) \sim \sum_{n=-\infty}^{\infty} c_n e^{inx}; \\ c_n = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-inx} dx. \end{cases}$$

Similarly, in the $L^2(-\infty, \infty)$ setting, if the Riesz basis $\{\psi_{j,k}(x)\}$ is an orthonormal (o.n.) basis of $L^2(-\infty, \infty)$; that is,

$$\langle \psi_{j,k}, \psi_{\ell,m} \rangle := \int_{-\infty}^{\infty} \psi_{j,k}(x) \overline{\psi_{\ell,m}(x)} dx = \delta_{j,\ell} \cdot \delta_{k,m}, \quad j, k, \ell, m \in \mathbb{Z},$$

then we also have

$$\begin{aligned} d_k^j &= \langle f, \psi_{j,k}(x) \rangle \\ &= \int_{-\infty}^{\infty} f(x) \overline{\psi_{j,k}(x)} dx. \end{aligned}$$

In general, if the Riesz basis $\{\psi_{j,k}(x)\}$ is not o.n., then we need the dual (or bi-orthogonal) basis $\{\psi^{j,k}(x)\}$, defined by

$$\langle \psi_{j,k}, \psi^{\ell,m} \rangle = \delta_{j,\ell} \delta_{k,m}, \quad j, k, \ell, m \in \mathbb{Z}.$$

Remark.

Although $\{\psi_{j,k}(x)\}$ is derived from a single function $\psi(x)$, the family $\{\psi^{\ell,m}(x)\}$ may not come from one single function $\tilde{\psi}$. If it does, then

$$\psi^{\ell,m}(x) \equiv \tilde{\psi}_{\ell,m}(x) := 2^{\ell/2} \tilde{\psi}(2^\ell x - m),$$

and we call $\tilde{\psi}$ the dual of ψ .

In general, if ψ has a dual $\tilde{\psi}$, then the series

$$f(x) = \sum_{j,k} d_k^j \psi_{j,k}(x)$$

is called a wavelet series. Hence, if $\{\psi_{j,k}\}$ is an o.n. basis of $L^2(-\infty, \infty)$, then since $\tilde{\psi} = \psi$, the above series expansion is a wavelet series.

The importance of the dual $\tilde{\psi}$ is that in the above wavelet series expansion of $f(x)$, we have

$$d_k^j \equiv (W_{\tilde{\psi}} f) \left(\frac{k}{2^j}, \frac{1}{2^j} \right),$$

where

$$(W_{\tilde{\psi}}f)(b, a) := \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(x) \overline{\tilde{\psi}\left(\frac{x-b}{a}\right)} dx$$

is called the **integral wavelet transform (IWT)** of $f(x)$ with respect to the mother (or basic) wavelet $\tilde{\psi}$. Hence, the wavelet series expansion of $f(x) \in L^2(-\infty, \infty)$ is given by

$$f(x) = \sum_{j,k \in \mathbb{Z}} \left\{ (W_{\tilde{\psi}}f) \left(\frac{k}{2^j}, \frac{1}{2^j} \right) \right\} \psi_{j,k}(x).$$

We conclude this discussion with the following observation.

Comparison of Fourier analysis with Wavelet analysis.

(a) In Fourier Analysis, we have two areas:

- (i) **Fourier series** and
- (ii) **Integral Fourier Transform**,

and they are not related.

(b) In Wavelet Analysis, we also have two areas:

- (i) **Wavelet series (WS)** and
- (ii) **Integral wavelet transform (IWT)**,

but these two areas are intimately related: the coefficients d_k^j of the WS of f in terms of a wavelet ψ are the IWT of f at $(\frac{k}{2^j}, \frac{1}{2^j})$, using the dual $\tilde{\psi}$ of ψ as the mother wavelet.

5. THE INTEGRAL WAVELET TRANSFORM AND TIME-FREQUENCY ANALYSIS.

The IWT of f

$$(W_{\psi}f)(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} f(t) \overline{\psi\left(\frac{t-b}{a}\right)} dt$$

may be considered as a "time-windowing process", where $f(t)$ is treated as a continuous-time signal. This window slides along the time-axis, while the width of the window is adjusted by the value of the scale $a > 0$. The smaller the value of a , the narrower the time-window; and consequently, a more accurate time-location is achieved.

By the Parseval identity, we have

$$(W_{\psi}f)(b, a) = \frac{\sqrt{a}}{2\pi} \int_{-\infty}^{\infty} \hat{f}(\omega) e^{ib\omega} \overline{\eta\left(a\left(\omega - \frac{\omega^*}{a}\right)\right)} d\omega,$$

where

$$\eta(\omega) := \hat{\psi}(\omega + \omega^*)$$

and ω^* is the "center" of $\hat{\psi}(\omega)$. Hence, the IWT also gives localized information on the spectral behavior $\hat{f}(\omega)$ of the signal $f(t)$. Again, the scale $a > 0$ adjusts the width of the frequency window.

If the frequency ω is set to be c/a for some calibration constant $c > 0$, then the time-scale plane becomes the time-frequency plane, and the "time-frequency window" narrows at high frequencies and widens at low frequencies.

6. WAVELET DECOMPOSITION AND RECONSTRUCTION ALGORITHMS.

Recall that

$$L^2(-\infty, \infty) \approx V_N = W_{N-1} \oplus \cdots \oplus W_{N-M} \oplus V_{N-M}.$$

Hence, any $f(x) \in L^2(-\infty, \infty)$ can be first approximated by an $f_N(x) \in V_N$ and then $f_N(x)$ is decomposed into a sum of its "wavelet components" $g_j(x)$ and its "blurred" version $f_{N-M}(x)$:

$$f \mapsto f_N = g_{N-1} + \cdots + g_{N-M} + f_{N-M}.$$

Write

$$\begin{cases} f_j(x) = \sum_{k=-\infty}^{\infty} c_k^j \phi(2^j x - k), \\ c^j = \{c_k^j\} \end{cases}$$

and

$$\begin{cases} g_j(x) = \sum_{k=-\infty}^{\infty} d_k^j \psi(2^j x - k), \\ d^j = \{d_k^j\}, \end{cases}$$

where ϕ generates V_0 (e.g. $\phi = N_1$).

Let $\{a_k, b_k\}$ be the "decomposition sequences" (relating $\phi(2x - \ell)$ with $\phi(x - k)$ and $\psi(x - k)$), and let $\{p_k, q_k\}$ be the "reconstruction sequences" (relating $\phi(x)$ with $\phi(2x - \ell)$, and $\psi(x)$ with $\phi(2x - \ell)$, respectively). Then the algorithms can be described as follows:

(i) **Decomposition algorithm**

$$\begin{cases} c_k^{j-1} = \sum_{\ell} b_{\ell-2k} c_{\ell}^j \\ d_k^{j-1} = \sum_{\ell} b_{\ell-2k} c_{\ell}^j \end{cases}$$

(Moving averaging followed by "downsampling")

$$\begin{array}{ccccccc} & & d^{N-1} & & d^{N-2} & & d^{N-M} \\ & \nearrow & & \nearrow & & \nearrow & \\ c^N & \longrightarrow & c^{N-1} & \longrightarrow & c^{N-2} & \longrightarrow & \cdots \longrightarrow c^{N-M} \end{array}$$

(ii) **Reconstruction algorithm**

$$c_k^j = \sum_l (p_{k-2l} c_l^{j-1} + q_{k-2l} d_l^{j-1})$$

("Upsampling" followed by moving averaging)

$$\begin{array}{ccccccc} d^{N-M} & & d^{N-M+1} & & & d^{N-1} & \\ & \searrow & & \searrow & & \searrow & \\ c^{N-M} & \longrightarrow & c^{N-M+1} & \longrightarrow & \dots & \longrightarrow & c^{N-1} \longrightarrow c^N \end{array}$$

7. INTEGRAL WAVELET TRANSFORM AS BANDPASS FILTERING.

Let $\psi(x)$ be a mother wavelet. For each (fixed) scale $a > 0$ (that decides the frequency range in the passband), set

$$h(t) = \frac{1}{\sqrt{a}} \overline{\psi\left(-\frac{t}{a}\right)}.$$

Then the IWT of f becomes

$$\begin{aligned} (W_\psi f)(b, a) &= \int_{-\infty}^{\infty} h(b-t) f(t) dt \\ &= (h * f)(b). \end{aligned}$$

That is, for fixed $a > 0$, W_ψ is a time-invariant linear filter with transfer function given by

$$H_a(\omega) = \hat{h}(\omega).$$

Remarks.

- (i) The bandpass filter has linear-phase if and only if $h(t)$ is symmetric; that is, there is some t_0 such that

$$\psi(t_0 + t) = \psi(t_0 - t).$$

- (ii) The bandpass filter has generalized linear-phase if and only if $h(t)$ is antisymmetric; that is, there is some t_0 such that

$$\psi(t_0 + t) = -\psi(t_0 - t).$$

- (iii) Linear-phase (or at least generalized linear-phase) $\psi(t)$ is essential for distortion-free filtering.

8. COMPARISON OF WAVELETS.

The Haar wavelet ψ_h does not give good frequency localization. So, in the following, we will assume that $\psi \neq \psi_h$.

(i) **Compactly supported o.n. wavelet ψ .**

(Note: $\{\psi_{j,k}\}$ is an o.n. basis of $L^2(-\infty, \infty)$.)

- All the decomposition and reconstruction sequences $\{a_n, b_n\}$ and $\{p_n, q_n\}$ are finite.
- ψ is **not** symmetric and **not** antisymmetric.
- Side loop/main loop ratio is fairly large.

(See Figure 1.)

(ii) **B-wavelets (of cardinal spline functions) ψ_m .**

- $W_n \perp W_j$, $n \neq j$.
- $\{a_n, b_n\}$ infinite, $\{p_n, q_n\}$ finite.
- $\psi_m(x)$ symmetric for even m , and antisymmetric for odd m (with respect to the center $\frac{2m-1}{2}$).
- Side loop/main loop ratio is smaller for $m > 2$.

(See Figures 2 and 3.)

(iii) **Symmetric or antisymmetric compactly supported ψ with finite decomposition and reconstruction sequences.**

- $W_n \not\perp W_j$.

9. APPLICATIONS.

Real-time wavelet decomposition algorithms can be implemented in parallel, separating the signal, image, etc., into "disjoint" frequency bands, with time (and/or space) localization (of high and low amplitudes) in each band. Furthermore, "wavelet-packet" tree algorithms can be implemented, even adaptively, to further decompose the components in the high-frequency ranges. Since the real-time wavelet reconstruction algorithms are equally efficient, it is easy to imagine that the list of potential applications of wavelets is endless. We only discuss a few examples here.

- (i) Analysis of transient signals
- (ii) Sonar applications
- (iii) Echo detection and cancellation
- (iv) Noise removal (e.g. acoustic pop noise)

(v) Image compression

Demonstrations of all these items were presented in the lecture.

Reference

1. C. K. Chui, *An Introduction to Wavelets*, Academic Press, Boston, 1992.

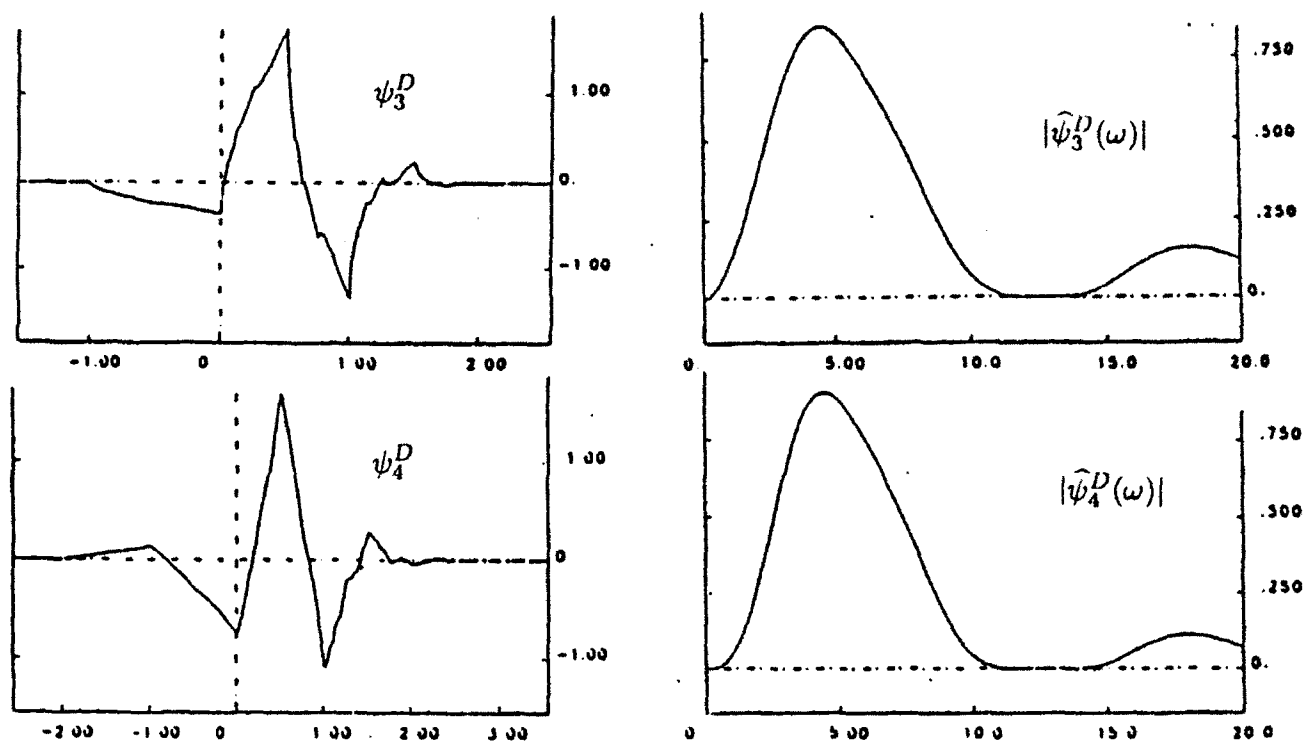


Figure 1. Daubechies wavelets and their filter characteristics.

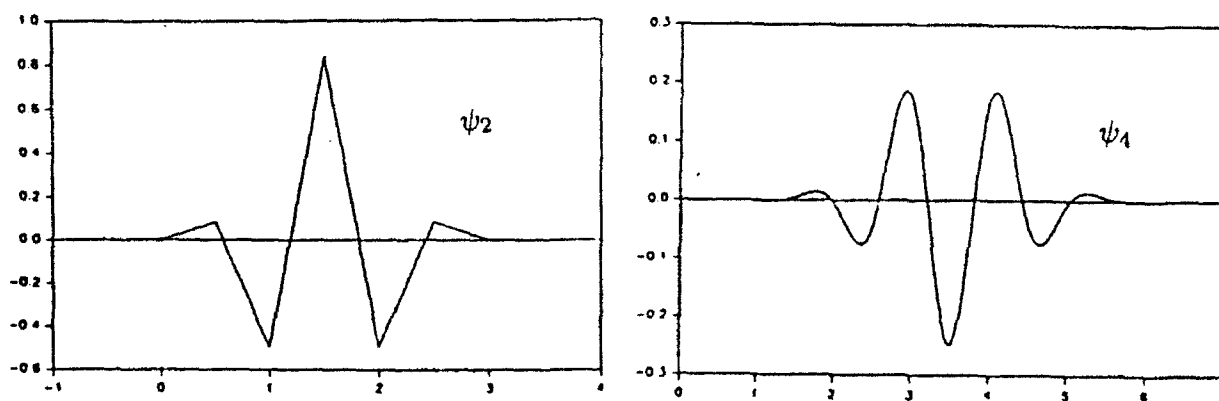


Figure 2. Linear and cubic B-wavelets.

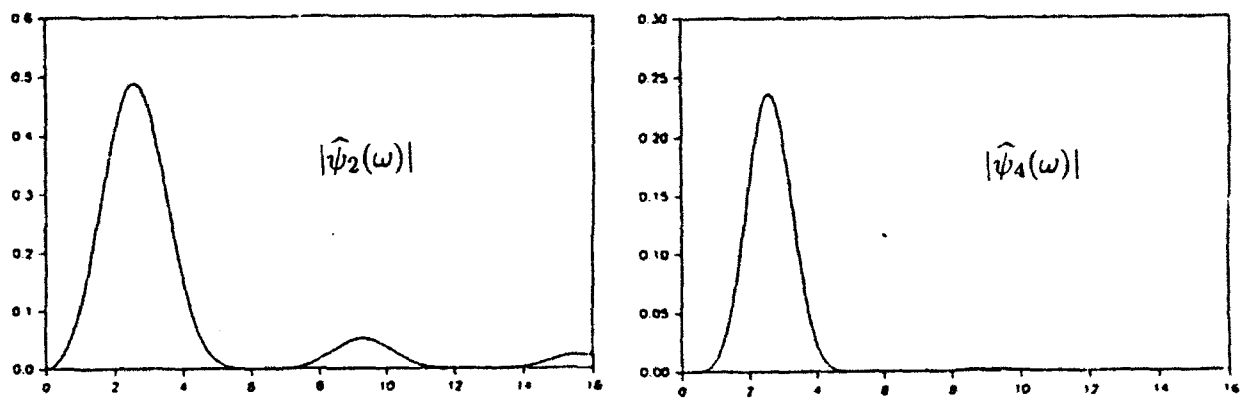


Figure 3. Filter characteristics of linear and cubic B-wavelets.

Design and Analysis of Scalable Parallel Algorithms. *

Vipin Kumar
Department of Computer Science,
University of Minnesota
Minneapolis, MN 55455
kumar@cs.umn.edu

Abstract

We have developed parallel algorithms and data structures for a variety of numeric and nonnumeric problems and analyzed their performance and scalability on various parallel architectures. Our analysis sheds light on what problems can be solved cost effectively on large-scale parallel computers. It also gives us insights into the best possible parallel architectures for solving various problems of practical interest. This paper presents an overview of our research. In particular, it summarizes results on development and analysis of parallel algorithms such as load balancing of unstructured tree computations, finite element method, sparse linear system solvers, backpropagation neural network learning algorithm and fast fourier transforms.

1 Introduction.

This research addresses the problem of exploiting the massive computation power of large scale parallel computers. It is possible to construct parallel processors containing 10's or 100's of thousands processors. The cost of these machines is comparable to that of large mainframes (\$ 1 to 20 Million), but they offer 100 to 10,000 fold more raw computing power. There are many important industrial/military problems that can become solvable with the 100 to 10,000 fold increment in computing power. It is not possible to exploit this massive power until scalable parallel algorithms are developed for the problems of interest. We are developing parallel algorithms for a variety of numeric and non-numeric problems and analyzing their performance and scalability on various parallel architectures. The scalability analysis is important because the hardware technology is changing rapidly (in terms of number of processors, computation and communication speeds and network technology) and experimental results for a problem on a specific architecture may become obsolete with any of these changes. This research will help answer the following questions: (i) what problems can be solved cost effectively on parallel computers? (ii) what are the best parallel algorithms and architectures for solving various problems of interest?

We have used the isoefficiency metric to analyze well known techniques for partitioning finite element meshes. Our analysis predicts the relative performance of these schemes on different parallel architectures and helps in determining more scalable schemes. We have analyzed the performance

*This work was supported by IST/SDIO through the Army Research Office grant # 28408-MA-SDI to the University of Minnesota and by the Army High Performance Computing Research Center at the University of Minnesota.

and scalability of the Preconditioned Conjugate Gradient Algorithm on parallel architectures such as mesh, hypercube and CM5TM¹ for different types of sparse matrices. We have predicted the relative performance of different mappings and the scalability of different preconditioners on large scale parallel computers. The analytical results have been verified through experimental implementations of these schemes on CM5. We have developed a new and highly scalable network partitioning method for mapping the Backpropagation Algorithm for parallel computers such as nCUBE2TM² and CM5. Even an unoptimized version of our hybrid parallel formulation on a 256 processor CM5 (without vector units) performs over 60 million weight changes per second (or over 180 million connections per second) for irregular networks. We have developed highly scalable methods for load balancing of unstructured tree computations on both MIMD and SIMD machines. We have studied the scalability and the cost-performance tradeoffs for the FFT algorithm on different architectures. This analysis highlights the dramatic impact of certain hardware parameters on the scalability and performance of the FFT algorithm. We have presented new parallel algorithms for matrix multiplication, and have analyzed the performance and scalability of these and existing algorithms for a variety of architectures.

In subsequent sections we provide an extended summary of our research results.

2 Analyzing the scalability of parallel algorithms and architectures.

At the given state of technology, it is possible to construct parallel computers employing hundreds of thousands of processors. Availability of such systems has fueled interest in investigating the performance of parallel computers containing a large number of processors. The scalability of a parallel algorithm on a parallel architecture is a measure of its capability to effectively utilize an increasing number of processors. Scalability analysis of a parallel algorithm-architecture combination can be used for a variety of purposes. It may be used to select the best algorithm-architecture combination for a problem under different constraints on the growth of the problem size and the number of processors. It may be used to predict the performance of a parallel algorithm and a parallel architecture for a large number of processors from the known performance on fewer processors. For a fixed problem size it may be used to determine the optimal number of processors to be used and the maximum possible speedup that can be obtained. The scalability analysis can also predict the impact of changing hardware technology on the performance and thus help design better parallel architectures for solving various problems.

A number of metrics for scalability analysis have been developed [21]. In [23], we presented isoefficiency as a metric for characterizing the scalability of parallel algorithm - architecture combinations. If a parallel algorithm is used to solve a problem instance of a fixed size, then the efficiency decreases as number of processors P increases. The reason is that the total overhead increases with P . For many parallel algorithms, for a fixed P , if the problem size W is increased, then the efficiency becomes higher (and approaches 1), because the total overhead grows slower than W . For these parallel algorithms, the efficiency can be maintained at a desired value (between 0 and 1) with increasing number of processors, provided the problem size is also increased. We call

¹CM5 is a trademark of the Thinking Machines Corporation.

²nCUBE2 is a trademark of the nCUBE corporation.

such algorithms **scalable** parallel algorithms.

Note that for a given parallel algorithm, for different parallel architectures, the problem size may have to increase at different rates w.r.t. P in order to maintain a fixed efficiency. The rate at which W is required to grow w.r.t. P to keep the efficiency fixed is essentially what determines the degree of scalability of the parallel algorithm for a specific architecture. For example, if W is required to grow exponentially w.r.t. P , then the algorithm-architecture combination is poorly scalable. The reason for this is that in this case it would be difficult to obtain good speedups on the architecture for a large number of processors, unless the problem size being solved is enormously large. On the other hand, if W needs to grow only linearly w.r.t. P , then the algorithm-architecture combination is highly scalable and can easily deliver linearly increasing speedups with increasing number of processors for reasonable problem sizes. If W needs to grow as $f(P)$ to maintain an efficiency E , then $f(P)$ is defined to be the **isoefficiency function** for efficiency E and the plot of $f(P)$ w.r.t. P is defined to be the **isoefficiency curve** for efficiency E . A lower bound on any isoefficiency function is that asymptotically, it should be at least linear. This follows from the fact that all problems have a sequential (*i.e.* non decomposable) component. Hence any algorithm which shows a linear isoefficiency on some architecture is optimally scalable on that architecture. Algorithms with isoefficiencies of $O(P \log^c P)$, for small constant c , are also reasonably optimal for practical purposes.

In [21], we critically assess the state of the art in the theory of scalability analysis, and to motivate further research on the development of new and more comprehensive analytical tools to study the scalability of parallel algorithms and architectures. We survey a number of techniques and formalisms that have been developed for studying the scalability issues, and discuss their interrelationships. We show some interesting relationships between the technique of isoefficiency analysis developed in [23] and many other methods for scalability analysis. We point out some of the weaknesses of the existing schemes, and discuss possible ways of extending them.

In [12], we study the impact of parallel processing overheads and the degree of concurrency of a parallel algorithm on the optimal number of processors to be used when the criterion for optimality is minimizing the parallel execution time. We also study a more general criterion of optimality and show how operating at the optimal point is equivalent to operating at a unique value of efficiency which is characteristic of the criterion of optimality and the properties of the parallel system under study. In this paper, we also show how the paper generalizes and/or extends earlier results of many other researchers.

3 Scalability Analysis of Load Balancing Algorithms.

Load balancing is perhaps the central aspect of parallel computing. Before a problem can be executed on a parallel computer, the work to be done has to be partitioned among different processors. Due to uneven processor utilization, load imbalance can cause poor efficiency. In [20], we have analyzed the problem of load balancing in multiprocessors for those parallel algorithms that have the following characteristics.

- The work available at any processor can be partitioned into independent work pieces as long as it is more than some non-decomposable unit.

- The cost of splitting and transferring work to another processor is not excessive. (i.e. the cost associated with transferring a piece of work is much less than the computation cost associated with it.)
- A reasonable work splitting mechanism is available; i.e., if work w at one processor is partitioned in 2 parts ψw and $(1 - \psi)w$, then $1 - \alpha > \psi > \alpha$, where α is an arbitrarily small constant.
- It is not possible (or is very difficult) to estimate the size of total work at a given processor.

Although, in such parallel algorithms, it is easy to partition the work into arbitrarily many parts, these parts can be of widely differing sizes. Hence after an initial distribution of work among P processors, some processors may run out of work much sooner than others; therefore a dynamic balancing of load is needed to transfer work from processors that have work to the ones that are idle. Since none of the processors (that have work) know how much work they have, load balancing schemes which require this knowledge (eg. [16, 18]) are not applicable. The performance of a load balancing scheme is dependent upon the degree of load balance achieved and the overheads due to load balancing.

Work created in the execution of many tree search algorithms used in artificial intelligence and operations research [22, 27] and many divide-and-conquer algorithms [15] satisfy all the requirements stated above. As an example, consider the problem of searching a state-space tree in depth-first fashion to find a solution. The state space tree can be easily split up into many parts and each part can be assigned to a different processor. Although it is usually possible to come up with a reasonable work splitting scheme [26], different parts can be of radically different sizes, and in general there is no way of estimating the size of a search tree.

A number of dynamic load balancing strategies that are applicable to problems with these characteristics have been developed. Many of these schemes have been experimentally tested on some physical parallel architectures. From these experimental results, it is difficult to ascertain relative merits of different schemes.

In [20, 9], we have been able to determine the most scalable load balancing schemes for different architectures such as hypercube, mesh and network of workstations. For each architecture, we have established lower bounds on the scalability of any possible load balancing scheme. We present the scalability analysis of a number of load balancing schemes that have not been analyzed before. From this we gain valuable insights into which schemes can be expected to perform better under what problem and architecture characteristics. For each of these architectures, we are able to determine near optimal load balancing schemes. In particular, some of the algorithms analyzed here for hypercubes are more scalable than those presented in [23]. Results obtained from implementation of these schemes in the context of the Tautology Verification problem on the Ncube/2TM multicomputer are used to validate theoretical results for the hypercube architecture.

In [17], we present new methods for load balancing of unstructured tree computations on large-scale SIMD machines, and analyze the scalability of these and existing schemes. An efficient formulation of tree search on a SIMD machine comprises of two major components: (i) a triggering mechanism, which determines when the search space redistribution must occur to balance search space over processors; and (ii) a scheme to redistribute the search space. We have devised a new redistribution mechanism and a new triggering mechanism. Either of these can be used in

conjunction with triggering and redistribution mechanisms developed by other researchers. We analyze the scalability of these mechanisms, and verify the results experimentally. The analysis and experiments show that our new load balancing methods are highly scalable on SIMD architectures. In particular, their scalability is no worse than that of the best load balancing schemes on MIMD architectures.

4 Scalability analysis of parallel formulations of the Fast Fourier Transform algorithm.

Fast Fourier Transform plays an important role in several scientific and technical applications. Some of the applications of the FFT algorithm include Time Series and Wave Analysis, solving Linear Partial Differential Equations, Convolution, Digital Signal Processing and Image Filtering, etc. Hence, there has been a great interest in implementing FFT on parallel computers. In [11], we analyze the scalability of a commonly used parallel formulation of FFT [30] on mesh and hypercube connected multicomputers. We also present experimental performance results on a 1024-processor Ncube/1 multicomputer to support analytical results.

The scalability analysis of FFT on hypercube provides several important insights. On the hypercube architecture, the parallel FFT algorithm can obtain linearly increasing speedup with respect to the number of processors with only a moderate increase in problem size. This is not surprising in the light of the fact that the FFT computation maps naturally to the hypercube architecture [28]. But there is a limit on the achievable efficiency which is determined by the ratio of CPU speed and communication bandwidth of the hypercube channels. This limit can be raised by increasing the bandwidth of the communication channels. Efficiencies higher than this limit can be obtained only if the problem size is increased very rapidly. The technology dependent features such as the communication bandwidth determine an upper-bound on the overall performance that one could obtain from a p -processor system for a given problem size. Thus if the processing speed of the CPU used in a p -processor hypercube is increased (and if these other factors are not changed), then the overall performance does not increase beyond a point. An interesting insight is that this upper-bound can be improved by either increasing the problem size exponentially or by improving the communication related parameters linearly.

From the scalability analysis, we found that the FFT algorithm cannot make efficient use of large-scale mesh architectures unless the communication bandwidth is increased as a function of the number of processors. If the width of inter-processor links is maintained as $O(\sqrt{p})$, where p is the number of processors on the mesh, then the scalability can be improved considerably. Addition of features such as cut-through-routing (also known as worm-hole routing) [3] to the mesh architecture improve the scalability of several parallel algorithms; *e.g.*, see [25]. But these features do not improve the overall scalability characteristics of the FFT algorithm on this architecture. We also show that if the cost of a communication network is proportional to the total number of communication links, then it is more cost-effective to implement the FFT algorithm on a hypercube rather than a mesh despite the fact that large scale meshes are cheaper to construct than large hypercubes.

We have used the single dimensional unordered radix-2 FFT algorithm for a major part of the analysis and for obtaining experimental results. This is the simplest form of FFT and not the most

efficient one. It is shown through similar analysis for multidimensional, ordered and higher radix algorithms, some of which are more efficient than the simple algorithm given here, that the nature of the results does not change [11].

5 Scalability analysis of parallel formulations of the Preconditioned Conjugate Gradient method.

In [13], we study performance and scalability of parallel formulations of the Preconditioned Conjugate Gradient (PCG) algorithm [7] for solving large sparse linear systems of equations of the form $Ax = b$, where A is a symmetric positive definite matrix. A linear system of equations is often *preconditioned* to accelerate the rate of convergence of the CG algorithm. In this paper, the use of two such preconditioning methods is considered - the diagonal preconditioner and that resulting from the Incomplete Cholesky (IC) factorization of the matrix of coefficients. Two different kinds of matrices are considered. First the scalability of the PCG algorithm with penta-diagonal matrices resulting from two dimensional square or rectangular finite difference grids with natural ordering of grid points is analyzed. Two commonly used schemes for mapping the data on the processors are compared and one is shown to be strictly better than the other one. These results are then extended to the matrices resulting from three dimensional finite difference grids. The second type of matrices that are studied are randomly sparse symmetric positive definite matrices. The analytical results are then verified through extensive experiments on the CM5 parallel computer. Apart from the basic questions answered by isoefficiency analysis, this analysis helps in answering a number of other questions, such as -

Which feature of the hardware should be improved for maximum returns in terms of performance per unit cost?

How does the Incomplete Cholesky (IC) preconditioner compare with a simple diagonal preconditioner in terms of parallel performance?

What kind of improvement in scalability can be achieved by re-ordering the sparse matrix?

Which parts of the algorithm dominate in terms of communication overheads and hence determine the overall parallel speedup and efficiency?

Although, we specifically deal with the Preconditioned CG algorithm only, the analysis pertaining to the diagonal preconditioner case applies to the non-preconditioned method also. In fact the results of the entire paper can be adapted for a number of iterative methods that use matrix-vector multiplication and vector inner product calculation as the basic operations in each iteration.

It is shown that for such matrices, the computation of vector inner products dominates the rest of the computation in terms of communication overheads. However, with a suitable mapping, the parallel formulation of the CG algorithm is highly scalable for such matrices on a machine like the CM5 whose fast control network practically eliminates the overheads due to inner product computation. A method of using the Incomplete Cholesky (IC) preconditioner is presented that leads to a further improvement in scalability on the CM5 by a constant factor. As a result, if enough processors are used, then a parallel implementation with the IC preconditioner may execute

faster than that with a simple diagonal preconditioner even if the latter executed faster in a serial implementation. For hepta-diagonal matrices resulting from three dimensional finite difference grids, the scalability is quite good on a hypercube or the CM5, but not as good on a 2-D mesh such as Intel Touchstone machine. In case of a random sparse matrix with a constant number of non-zero elements in each row, the parallel formulation of the CG method is unscalable on any parallel architecture. But the parallel system can be rendered scalable either if, after reordering, the non-zero elements of the $N \times N$ matrix can be confined in a band whose width is $O(N^y)$ for any $y < 1$, or if the number of non-zero elements per row increases as N^x for any $x > 0$. The scalability increases as the number of non-zero elements per row is increased and/or the width of the band containing these elements is reduced. For random sparse matrices, the scalability is asymptotically the same for all architectures.

6 Scalability analysis of parallel algorithms for matrix multiplication.

Matrix multiplication is widely used in a variety of applications and is often one of the core components of many scientific computations. Since the matrix multiplication algorithm is highly computation intensive, there has been a great deal of interest in developing parallel formulations of this algorithm and testing its performance on various parallel architectures.

Some of the early parallel formulations of matrix multiplication were developed by Canon [2], Dekel, Nassimi and Sahni [4], and Fox *et. al.* [5]. Variants and improvements of these algorithms have been presented in [1, 14]. In particular, Berntsen [1] presents an algorithm which has a strictly smaller communication overhead than Canon's algorithm, but has a smaller degree of concurrency. Ho and Johnsson [14] present another variant of Canon's algorithm for a hypercube which permits communication on all channels simultaneously. This algorithm too, while reducing communication, also reduces the degree of concurrency.

In [10], we use the isoefficiency metric [23] to analyze the scalability of a number of parallel formulations of the matrix multiplication algorithm for a wrap-around mesh, hypercube and related architectures. We analyze the performance of various parallel formulations of the matrix multiplication algorithm for different matrix sizes and number of processors, and predict the conditions under which each formulation is better than the others. We present a new parallel algorithm for the hypercube and related architectures that performs better than any of the schemes described in the literature so far for a wide range of matrix sizes and number of processors. The superior performance and the analytical scalability expressions for this algorithm are verified through experiments on the CM5 parallel computer for upto 512 processors. We show that special hardware permitting simultaneous communication on all the ports of the processors does not improve the overall scalability of the matrix multiplication algorithms on a hypercube. We also discuss the dependence of scalability of parallel matrix multiplication algorithms on technology dependent factors such as communication and computation speeds and show that under certain conditions, it may be better to have a parallel computer with k -fold as many processors rather than one with the same number of processors, each k -fold as fast.

7 Scalability analysis of partitioning techniques for finite element graphs.

Parallel formulations of finite element techniques require a mapping of the elements onto processors, and the performance of the overall formulation is a very sensitive function of this mapping. Any mapping of elements to processors must try to satisfy the following criteria:

1. The ratio of computation to communication associated with elements on a processor should be maximized.
2. Locality of communication should be preserved.
3. The computational load should be balanced to the extent possible.

These conditions represent the classical communication - load imbalance tradeoffs. Optimizing one of these criterion leads to a deterioration with respect to one or more of the other criteria. The mapping problem in its optimal form is known to be NP complete even for simple models of computation and communication costs [6]. Hence, a number of heuristic approaches have been presented to derive reasonable suboptimal partitions in a reasonable amount of time. All of these try to balance the various tradeoffs mentioned. Most of these schemes have been evaluated only on specific parallel computers for certain problems.

In [8], we perform scalability analysis, using the Isoefficiency metric [24, 25], of three partitioning algorithms, namely, striped partitioning, binary decomposition, and scattered decomposition. This helps us establish the relative performance of these schemes over a range of processors, and the effect of communication related parameters on the performance of these schemes. We also relate the performance of each of these schemes to the various problem characteristics such as mesh geometry and density. The theoretical results are verified through simulations.

8 Scalable Parallel Formulations of the Backpropagation Algorithm.

The Backpropagation algorithm (BP)[29] is one of the most popular neural network learning algorithms. It has been used in a large number of applications.

BP can be parallelized either by network partitioning or by pattern partitioning. In network partitioning schemes, nodes and weights of the neural network are partitioned among different processors and thus the computations of node activations, node errors and weight changes are parallelized. In pattern partitioning, individual weight changes due to various learning patterns are computed concurrently. Pattern partitioning and network partitioning can also be combined to form hybrid schemes. Several machine architectures including linear arrays, meshes and hypercubes have been explored to implement parallel BP.

In [19], we present a new technique for mapping the backpropagation algorithm on hypercubes and related architectures. A key component of this technique is a network partitioning scheme which is called *checkerboarding*. The major communication intensive operation in the commonly used vertical sectioning scheme is the all-to-all broadcast. In this operation, each of the P processor has to broadcast its local m units of information to all other processors. This takes $O(mP)$ time

on linear array, mesh as well as hypercube. Hence, the vertical sectioning scheme performs equally well (or equally poorly) on all these architectures.

The checkerboarding scheme allows us to replace the all-to-all broadcast operation by concurrent non-interfering single source broadcasts, which are much faster on hypercubes. Furthermore, our method can use a larger number of processors without incurring higher communication costs than the vertical sectioning scheme, and shows strictly better performance on hypercubes with more than 16 processors. Our scheme also performs better than pattern partitioning scheme for a large class of problems. In addition, our scheme can be combined with the pattern partitioning scheme to form a hybrid scheme which performs better than either one for a wider variety of cases. Although the scheme is natural for the hypercube architecture (e.g., nCUBE, Intel iPSC^{TM3}), it is equally suitable for Fat-tree based architectures such as CM5. Our scheme is applicable only to fully connected networks; i.e., each node in a layer (except the output layer) is connected to all nodes in the next layer. The number of nodes in each layer can be the same (uniform network) or different (non-uniform network).

Experimental results on nCUBE2 and CM5 show that our scheme performs better than the other schemes for both uniform and non-uniform networks. Furthermore, it provides very high overall performance on existing commercial parallel computers. For example, an unoptimized version of our hybrid parallel formulation on a 256 processor CM5 (without vector units) performs over 50 million weight changes per second (or over 160 million connections per second) for non-uniform networks. With the upgrade of processors on the CM5, this performance is expected to improve by 1 or 2 orders of magnitude.

9 Concluding Remarks.

Development of efficient parallel algorithms and the understanding of their scalability for the problems mentioned above is very useful for both military and industrial applications. Finite element and finite difference methods are widely used in modeling fluid flows, fluid dynamics and many other applications. Unstructured tree search algorithms such as Branch and Bound are used to solve combinatorial optimization problems such as resource allocation, logistics and transportation problems, etc. Dynamic load balancing techniques play a very important role in various problems including discrete event modeling such as battlefield simulations. Neural Network learning algorithms such as Backpropagation are important components of target recognition systems. The FFT algorithm is an integral part of many applications that involve signal processing. Our research has led to a better understanding of the relative merits of various parallel formulations of different algorithms and development of new and more scalable formulations for many of these problems.

References

- [1] Jarle Berntsen. Communication efficient matrix multiplication on hypercubes. *Parallel Computing*, 12:335 - 342, 1989.
- [2] L. E. Cannon. A cellular computer to implement the Kalman Filter Algorithm. Technical Report PhD Thesis, Montana State University, 1969.

³iPSC is a trademark of the Intel Scientific Computers

- [3] William Dally. *A VLSI Architecture for Concurrent Data Structures*. Kluwer Academic Publ, Boston, Massachusetts, 1987.
- [4] Eliezer Dekel, David Nassimi, and Sartaj Sahni. Parallel matrix and graph algorithms. *SIAM Journal of Computing*, 10:657 - 673, 1981.
- [5] G.C. Fox, S.W. Otto, and A.J.G. Hey. Matrix algorithms on a hypercube i: Matrix multiplication. *Parallel Computing*, 4:17-31, 1987.
- [6] M. Garey and D.S. Johnson. *Computers and Intractability*. Freeman, San Francisco, 1979.
- [7] Gene H. Golub and Charles F. Van Loan. *Matrix Computations; second edition*. The John Hopkins University Press, 1989.
- [8] Ananth Grama and Vipin Kumar. Scalability analysis of partitioning strategies for finite element graphs. Technical report, Tech Report , Computer Science Department, University of Minnesota, 1992.
- [9] Ananth Grama, Vipin Kumar, and V. Nageshwara Rao. Experimental evaluation of load balancing techniques for the hypercube. In *Proceedings of the Parallel Computing 91 Conference*, 1991.
- [10] Anshul Gupta and Vipin Kumar. On the scalability of Matrix Multiplication Algorithms on parallel computers. Technical Report TR 91-54, Computer Science Department, University of Minnesota, Minneapolis, MN 55455, 1991.
- [11] Anshul Gupta and Vipin Kumar. On the scalability of FFT on parallel computers. *IEEE Transactions on Parallel and Distributed Systems*, 1993 (to appear). available as a technical report TR 90-20, Computer Science Department, University of Minnesota.
- [12] Anshul Gupta and Vipin Kumar. Analyzing performance of large scale parallel systems. Technical report, Computer Science Department, University of Minnesota, Minneapolis, MN - 55455, May 1991.
- [13] Anshul Gupta, Vipin Kumar, and Ahmed Sameh. Performance and scalability of conjugate gradient methods on parallel computers. Technical report, University of Minnesota, 1992.
- [14] Ching-Tien Ho, S. Lennart Johnsson, and Alan Edelman. Matrix multiplication on hypercubes using full bandwidth and constant storage. In *Proceedings of the 1991 International Conference on Parallel Processing*, pages 447 - 451, 1991.
- [15] Ellis Horowitz and Sartaj Sahni. *Fundamentals of Computer Algorithms*. Computer Science Press, Rockville, Maryland, 1978.
- [16] L. V. Kale. Comparing the performance of two dynamic load distribution methods. In *Proceedings of International conference on Parallel Processing*, pages 8-12, 1988.
- [17] George Karypis and Vipin Kumar. Unstructured Tree Search on SIMD Parallel Computers. Technical Report 92-21, Computer Science Department, University of Minnesota, 1992. A short version of this paper appears in the Proceedings of Supercomputing 1992 Conference, November 1992.

- [18] R. Keller and F. Lin. Simulated performance of a reduction based multiprocessor. *IEEE Computers*, July 1984 1984.
- [19] V. Kumar, S. Shekhar, and M. B. Amin. A highly parallel formulation of backpropagation on hypercubes: A summary of results. November 1992. Extended version available as TR92-54 from Computer science department, University of Minnesota, Minneapolis, MN 55455.
- [20] Vipin Kumar, Ananth Grama, and V. Nageshwara Rao. Scalable load balancing techniques for parallel computers. Technical report, Tech Report 91-55, Computer Science Department, University of Minnesota, 1991.
- [21] Vipin Kumar and Anshul Gupta. Analyzing scalability of parallel algorithms and architectures. Technical report, TR-91-18, Computer Science Department, University of Minnesota, June 1991. A short version of the paper appears in the Proceedings of the 1991 International Conference on Supercomputing, Germany, and as an invited paper in the Proc. of 29th Annual Allerton Conference on Communication, Control and Computing, Urbana, IL, October 1991.
- [22] Vipin Kumar, Dana Nau, and Laveen Kanal. General branch-and-bound formulation for and/or graph and game tree search. In Laveen Kanal and Vipin Kumar, editors, *Search in Artificial Intelligence*. Springer-Verlag, New York, 1988.
- [23] Vipin Kumar and V. Nageshwara Rao. Parallel depth-first search, part II: Analysis. *International Journal of Parallel Programming*, 16 (6):501-519, 1987.
- [24] Vipin Kumar and V. Nageshwara Rao. Load balancing on the hypercube architecture. In *Proceedings of the 1989 Conference on Hypercubes, Concurrent Computers and Applications*, pages 603-608, 1989.
- [25] Vipin Kumar and Vineet Singh. Scalability of Parallel Algorithms for the All-Pairs Shortest Path Problem. *Journal of Parallel and Distributed Processing (special issue on massively parallel computation)*, 13(2):124-138, October 1991. A short version appears in the Proceedings of the International Conference on Parallel Processing, 1990.
- [26] V. Nageshwara Rao and Vipin Kumar. Parallel depth-first search, part I: Implementation. *International Journal of Parallel Programming*, 16 (6):479-499, 1987.
- [27] Judea Pearl. *Heuristics - Intelligent Search Strategies for Computer Problem Solving*. Addison-Wesley, Reading, MA, 1984.
- [28] M. C. Pease. The indirect binary n-cube microprocessor array. *IEEE Transactions on Computers*, 26:458 - 473, 1977.
- [29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. *Learning internal representations by error propagation*, chapter 8. MIT Press, 1986.
- [30] P. N. Swartztrauber. Multiprocessor ffts. *Parallel Computing*, 5:197-210, 1987.

DYNAMICAL SYSTEMS IN ASYMMETRIC SPACE AND TIME

Richard A. Weiss

U.S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. The basic equations of kinematics and dynamics are derived for particles moving in space and time with broken internal symmetries. Within this formalism the space and time coordinates, velocities and accelerations of particles must be described as complex numbers having magnitudes and phase angles which can be variables. Incoherent spacetime is associated with changes in the magnitudes of the space and time coordinates while coherent spacetime is associated with the rotation of the coordinates in an internal space. Newton's law of dynamics is formulated for broken symmetry spacetime in four ways which allow the description of four limiting kinematic states of motion: incoherent space and incoherent time, coherent space and incoherent time, incoherent space and coherent time, and finally coherent space and coherent time. The relationship between the measured velocity and acceleration of particles and the conventionally calculated values of velocity and acceleration is determined for spacetime with broken internal symmetries. An application to the internal space and time motions of the harmonic oscillator is presented as an elementary example of the theoretical formalism.

1. INTRODUCTION. It has been suggested that the thermodynamic functions such as pressure, internal energy and entropy are skewed in an internal space, and must be represented by complex numbers with internal phase angles.¹ This conclusion follows from the assumption of the validity of a relativistic trace equation which introduces the effects of the Grüneisen function and the bulk modulus on the relativistic state equation for a material system.^{1,2} The broken symmetry nature of the pressure combined with Euler's equations of motion for fluids suggest that the space and time coordinates within matter have internal phase angles and must be represented as complex numbers in an internal space.¹ It then follows naturally that the kinematic quantities such as particle velocity, momentum, acceleration and energy also have internal phase angles. Dynamical quantities such as force, Lagrangian, Hamiltonian and action variables must also have broken internal symmetries and be represented by complex numbers in an internal space. In this way the broken symmetries of matter and spacetime at a macroscopic scale of pressure and energy density require the microscopic formalism of particle dynamics to also have internal phase angles for the dynamical variables. The reverse argument can also be made that the broken symmetry of the microscopic kinematic and dynamical variables leads to internal phase angles being associated with pressure and internal energy. The internal structure of space and time affects the measured kinematic and dynamical quantities of mechanics. In this way mechanics, the oldest discipline of physics, can be related to the internal structure of space and time which on a macroscopic scale is determined mainly by gravity. On a microscopic scale the possible internal motions of space and time can be related to the structure of matter. For instance, in high- T_c superconducting materials it is thought that both space and time are coherent and that the Cooper electron pairs move about

each other by internal spacetime motions. In this way the factor $6/\pi$ enters the calculation of the normalized superconducting energy gap and gives rise to it's possible high values compared to the corresponding values predicted by standard BCS superconductivity theory.³

The space and time coordinates must be written as complex numbers in internal space as follows¹

$$\bar{v} = v \exp(j\theta_v) \quad (1)$$

$$\bar{t} = t \exp(j\theta_t^v) \quad (2)$$

where $v = x, y, z$ for cartesian coordinates; r, ϕ, z for cylindrical polar coordinates; and ρ, ϕ, ψ for spherical polar coordinates. Strictly speaking the internal phase angle of time θ_t^v is associated with each space coordinate so that for the three elementary coordinate systems the internal phase angles of time are:

$$\theta_t^x, \theta_t^y, \theta_t^z \quad \theta_t^r, \theta_t^\phi, \theta_t^z \quad \theta_t^\rho, \theta_t^\phi, \theta_t^\psi \quad (3)$$

The following quantities often appear in the calculations involving partially coherent broken symmetry space and time³

$$\tan \beta_{vv} = v \partial \theta_v / \partial v \quad (4)$$

$$\tan \beta_{tt}^v = t \partial \theta_t^v / \partial t \quad (5)$$

as for example in the following differentials

$$\begin{aligned} d\bar{v} &= \sec \beta_{vv} dv \exp[j(\theta_v + \beta_{vv})] \\ &= \csc \beta_{vv} v d\theta_v \exp[j(\theta_v + \beta_{vv})] \end{aligned} \quad (6)$$

$$\begin{aligned} d\bar{t} &= \sec \beta_{tt}^v dt \exp[j(\theta_t^v + \beta_{tt}^v)] \\ &= \csc \beta_{tt}^v t d\theta_t^v \exp[j(\theta_t^v + \beta_{tt}^v)] \end{aligned} \quad (7)$$

For incoherent spacetime $\beta_{vv} = 0$ and $\beta_{tt}^v = 0$ while for coherent spacetime $\beta_{vv} = \pi/2$ and $\beta_{tt}^v = \pi/2$. The measured coordinates of space and time are given by¹

$$v_m = v \cos \theta_v \quad t_m^v = t \cos \theta_t^v \quad (8)$$

The complex number coordinate speed is obtained from equations (6) and (7) as

$$\bar{v}_v = v_v \exp(j\theta_{vv}) = d\bar{v}/d\bar{t} \quad (9)$$

where

$$v_v = \cos \beta_{tt}^v \sec \beta_{vv} dv/dt \quad (10)$$

$$= \cos \beta_{tt}^v \csc \beta_{vv} v d\theta_v/dt \quad (11)$$

$$= \sin \beta_{tt}^v \sec \beta_{vv} t^{-1} dv/d\theta_t^v \quad (12)$$

$$= \sin \beta_{tt}^v \csc \beta_{vv} v/t d\theta_v/d\theta_t^v \quad (13)$$

where

$$\theta_{vv} = \theta_v + \beta_{vv} - \theta_t^v - \beta_{tt}^v \quad (14)$$

and where $v = x, y, z; r, \phi, z$; or ρ, ϕ, ψ . The particle velocity, momentum and energy are written as complex numbers in internal space as follows¹

$$\bar{v}_v = v_v \exp(j\theta_{vv}) \quad \bar{p}_v = p_v \exp(j\theta_{pv}) \quad \bar{E} = E \exp(j\theta_{Ev}) \quad (15)$$

and the corresponding measured quantities are

$$v_{vm} = v_v \cos \theta_{vv} \quad p_{vm} = p_v \cos \theta_{pv} \quad E_m = E \cos \theta_E \quad (16)$$

For the case of harmonic motion the space coordinates are complex numbers in both the internal and the external spaces, and the coordinate magnitudes that appear in equation (1) are written for cartesian coordinates as

$$x = x_\omega \exp(i\omega t) \quad y = y_\omega \exp(i\omega t) \quad z = z_\omega \exp(i\omega t) \quad (17)$$

so that the full expression for the complex number cartesian coordinates represented by equation (1) are

$$\bar{x} = x_\omega \exp(i\omega t + j\theta_x) \quad (18)$$

$$\bar{y} = y_\omega \exp(i\omega t + j\theta_y) \quad (19)$$

$$\bar{z} = z_\omega \exp(i\omega t + j\theta_z) \quad (20)$$

or in general for harmonic motion

$$\bar{v} = v \exp(j\theta_v) \quad v = v_\omega \exp(i\omega t) \quad (21)$$

where $v = x, y, z$ or r, ϕ, z or ρ, ϕ, ψ . From equations (4) and (21) for harmonic motion it follows that

$$\tan \beta_{vv} = -1/\omega \partial \theta_v / \partial t \quad (22)$$

This suggests that for harmonic motion in external space the angle β_{vv} is an imaginary number in external space given by

$$\beta_{vv} = -iQ_v \quad (23)$$

where Q_v = real number. Then

$$\tan \beta_{vv} = -i \tanh Q_v \quad (24)$$

which gives for harmonic motion in external space

$$\tanh Q_v = \omega^{-1} \partial \theta_v / \partial t \quad (25)$$

$$Q_v = \tanh^{-1} (\omega^{-1} \partial \theta_v / \partial t) \quad (26)$$

$$\beta_{vv} = -i \tanh^{-1} (\omega^{-1} \partial \theta_v / \partial t) \quad (27)$$

for $v = x, y, z$ or r, ϕ, z or ρ, ϕ, ψ .

The broken symmetry sine and cosine functions are written as¹

$$\sin \bar{\psi} = S_\psi \exp(j\theta_{s\psi}) \quad \cos \bar{\psi} = C_\psi \exp(-j\theta_{c\psi}) \quad (28)$$

$$S_\psi = [\sin^2(\psi \cos \theta_\psi) + \sinh^2(\psi \sin \theta_\psi)]^{1/2} \quad (29)$$

$$C_\psi = [\cos^2(\psi \cos \theta_\psi) + \sinh^2(\psi \sin \theta_\psi)]^{1/2} \quad (30)$$

$$\tan \theta_{s\psi} = \cot(\psi \cos \theta_\psi) \tanh(\psi \sin \theta_\psi) \quad (31)$$

$$\tan \theta_{c\psi} = \tan(\psi \cos \theta_\psi) \tanh(\psi \sin \theta_\psi) \quad (32)$$

For small values of angle magnitude ψ

$$S_\psi \sim \psi \quad \theta_{s\psi} \sim \theta_\psi \quad (33)$$

which are useful expressions for studying the motion of a pendulum.

This simple introduction to broken spacetime symmetry is sufficient to proceed to the study of dynamical systems in asymmetric spacetime. This paper examines the effects of broken spacetime symmetries on the basic laws and formalisms of kinematics and dynamics. Only an elementary development of the mechanics of a single particle is given. Specifically, Section 2 considers kinematics in asymmetric spacetime, and Section 3 examines Newton's laws of motion in spacetime with broken internal symmetries.

2. KINEMATICS IN ASYMMETRIC SPACETIME. This section develops the basic expressions for particle speed and acceleration in spacetime with broken internal symmetries. The concept of motion in totally coherent spacetime is introduced. The connections are made between the measured kinematical quantities of partially coherent spacetime and the conventionally calculated kinematical

quantities of incoherent spacetime.

A. Particle Speed.

The speed of a particle in broken symmetry spacetime is obtained from equations (9) through (14)

$$\bar{v}_x = v_x \exp(j\theta_{vx}) = d\bar{x}/d\bar{t} \quad (34)$$

where

$$v_x = \cos \beta_{tt}^x \sec \beta_{xx} dx/dt \quad (35)$$

$$= \cos \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt \quad (36)$$

$$= \sin \beta_{tt}^x \sec \beta_{xx} t^{-1} dx/d\theta_t^x \quad (37)$$

$$= \sin \beta_{tt}^x \csc \beta_{xx} x/t d\theta_x/d\theta_t^x \quad (38)$$

and

$$\theta_{vx} = \theta_x + \beta_{xx} - \theta_t^x - \beta_{tt}^x \quad (39)$$

where β_{xx} and β_{tt}^x are given by equations (4) and (5) respectively. The measured particle speed is given by equations (16) and (35) through (39) as

$$v_{mx} = v_x \cos \theta_{vx} \quad (40)$$

The single particle momentum is then given by

$$\bar{p}_x = p_x \exp(j\theta_{px}) \quad p_x = mv_x \quad \theta_{px} = \theta_{vx} \quad (41)$$

Equations similar to (34) through (39) can be developed for the y and z coordinates.

B. Particle Acceleration.

For a broken symmetry spacetime the particle acceleration is given by equations (34) through (39) as

$$\bar{a}_x = a_x \exp(j\theta_{ax}) = d\bar{v}_x/d\bar{t} = d^2\bar{x}/d\bar{t}^2 \quad (42)$$

From Newton's law of motion written as

$$\bar{F}_x = F_x \exp(j\theta_{Fx}) = m\bar{a}_x \quad (43)$$

it follows that

$$F_x = ma_x \quad \theta_{Fx} = \theta_{ax} \quad (44)$$

The measured acceleration is given by

$$a_{mx} = a_x \cos \theta_{ax} \quad (45)$$

The measured force is given by equations (43) and (44) as

$$F_{mx} = F_x \cos \theta_{Fx} \quad (46)$$

Combining equations (44) through (46) gives

$$F_{mx} = ma_{mx} \quad (47)$$

and therefore the measured acceleration is determined from the measured force by Newton's law of motion. The values of the acceleration magnitude a_x and acceleration internal phase angle θ_{ax} will now be calculated for several cases of interest.

Case a. Incoherent Space and Incoherent Time.

A general expression for the acceleration is developed that can be used to deduce the limiting case of incoherent space and incoherent time which is described by

$$\theta_x = 0 \quad \beta_{xx} = 0 \quad \theta_t = 0 \quad \beta_{tt}^x = 0 \quad (48)$$

The appropriate expressions for the acceleration magnitude and internal phase angle are deduced from equation (42) to be

$$a_x = \sec \beta_{vxvx} \cos \beta_{tt}^x dv_x/dt \quad (49)$$

$$\theta_{ax} = \theta_{vx} + \beta_{vxvx} - \theta_t^x - \beta_{tt}^x \quad (50)$$

where

$$\tan \beta_{vxvx} = v_x \partial \theta_{vx} / \partial v_x \quad (51)$$

Combining equations (35) and (49) gives

$$a_x = \cos \beta_{tt}^x \sec \beta_{vxvx} d/dt (\cos \beta_{tt}^x \sec \beta_{xx} dx/dt) \quad (52)$$

$$\sim \cos^2 \beta_{tt}^x \sec \beta_{vxvx} \sec \beta_{xx} d^2x/dt^2 \quad (53)$$

$$\sim \cos^2 \beta_{tt}^x \sec^2 \beta_{xx} d^2x/dt^2 \quad (54)$$

Combining equations (39) and (50) gives

$$\theta_{ax} = \theta_x + \beta_{xx} + \beta_{vxvx} - 2(\theta_t^x + \beta_{tt}^x) \quad (55)$$

When the conditions in equation (48) are valid the expressions in equations (52) through (54) reduce to the standard case of incoherent space and incoherent time.

Case b. Coherent Space and Incoherent Time.

General expressions for the acceleration are now deduced which can be used to make the transition to the case of coherent space and incoherent time which is described by

$$\beta_{xx} = \pi/2 \quad \theta_t^x = 0 \quad \beta_{tt}^x = 0 \quad (56)$$

The appropriate general expression for the acceleration magnitude and internal phase angle for this case is obtained from equations (42) and (55) to be

$$a_x = \csc \beta_{v xv x} \cos \beta_{tt}^x v_x d\theta_{vx}/dt \quad (57)$$

$$\theta_{ax} = \theta_x + \beta_{xx} + \beta_{v xv x} - 2(\theta_t^x + \beta_{tt}^x) \quad (58)$$

Combining equations (36) and (57) gives

$$a_x = \csc \beta_{v xv x} \cos^2 \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt d\theta_{vx}/dt \quad (59)$$

where from equation (39)

$$d\theta_{vx}/dt = d/dt(\theta_x + \beta_{xx} - \theta_t^x - \beta_{tt}^x) \quad (60)$$

For the case at hand it is convenient to write the acceleration in equation (42) as

$$\begin{aligned} \bar{a}_x &= a_x \exp(j\theta_{ax}) = a_x^+ \exp(j\theta_{ax}^+) \\ &= d\bar{v}_x/d\bar{t} = d^2\bar{x}/d\bar{t}^2 \end{aligned} \quad (61)$$

where

$$a_x^+ = -a_x \quad (62)$$

$$\theta_{ax}^+ = \theta_{ax} - \pi \quad (63)$$

so that

$$a_x^+ = -\csc \beta_{v xv x} \cos \beta_{tt}^x v_x d\theta_{vx}/dt \quad (64)$$

$$= -\csc \beta_{v xv x} \cos^2 \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt d\theta_{vx}/dt \quad (65)$$

$$\theta_{ax}^+ = \theta_x + \beta_{xx} + \beta_{v xv x} - 2(\theta_t^x + \beta_{tt}^x) - \pi \quad (66)$$

which is an equivalent description of the acceleration.

For the special case of coherent space and incoherent time, equations (56), (65) and (66) give

$$a_x^{ci+} = - \csc \beta_{vxvx}^{ci} x (d\theta_x/dt)^2 \quad (67)$$

$$\theta_{ax}^{ci+} = \theta_x + \beta_{vxvx}^{ci} - \pi/2 \quad (68)$$

where from equations (36), (39), (51) and (56) it follows that

$$v_x^{ci} = x d\theta_x/dt \quad (69)$$

$$\theta_{vx}^{ci} = \theta_x + \pi/2 \quad (70)$$

$$\tan \beta_{vxvx}^{ci} = E_{xt}^{ci}/F_{xt}^{ci} \quad (71)$$

where

$$E_{xt}^{ci} = (d\theta_x/dt)^2 \quad E_{xt}^{ci} \geq 0 \quad (72)$$

$$F_{xt}^{ci} = d^2\theta_x/dt^2 \quad F_{xt}^{ci} \leq 0 \quad (73)$$

Equation (71) then gives

$$\csc \beta_{vxvx}^{ci} = [(E_{xt}^{ci})^2 + (F_{xt}^{ci})^2]^{1/2}/E_{xt}^{ci} \quad (74)$$

Because $F_{xt}^{ci} \leq 0$ it follows from equation (71) that

$$\beta_{vxvx}^{ci} = \pi/2 + \kappa_{xt} \quad (75)$$

where $\kappa_{xt} \geq 0$ is a small number which is also given by

$$\tan \kappa_{xt} = |F_{xt}^{ci}|/E_{xt}^{ci} \quad (76)$$

Combining equations (67), (68), (74) and (75) gives

$$a_x^{ci+} = - x [(E_{xt}^{ci})^2 + (F_{xt}^{ci})^2]^{1/2} \quad (77)$$

$$\theta_{ax}^{ci+} = \theta_x + \kappa_{xt} \quad (78)$$

The angle θ_{ax}^{ci+} is generally a small number. Clearly the acceleration in this case is directed opposite to the displacement x .

Case c. Incoherent Space and Coherent Time.

An expression for the acceleration is now derived from which the limiting case of incoherent space and coherent time can be obtained. This limiting case is described by

$$\theta_x = 0 \quad \beta_{xx} = 0 \quad \beta_{tt}^x = \pi/2 \quad (79)$$

The required general expression for the acceleration magnitude and acceleration phase angle is obtained from equations (42) and (55) as

$$a_x = \sec \beta_{v xv x} \sin \beta_{tt}^x t^{-1} dv_x / d\theta_t^x \quad (80)$$

$$\theta_{ax} = \theta_x + \beta_{v xv x} + \beta_{xx} - 2(\theta_t^x + \beta_{tt}^x) \quad (81)$$

Combining equations (37) and (80) gives

$$a_x = \sec \beta_{v xv x} \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \sec \beta_{xx} t^{-1} dx / d\theta_t^x) \quad (82)$$

$$\sim \sec \beta_{v xv x} \sin^2 \beta_{tt}^x \sec \beta_{xx} t^{-2} d^2 x / d\theta_t^{x2} \quad (83)$$

For this case it is convenient to rewrite the acceleration equation (42) in the following form

$$\bar{a}_x = a_x \exp(j\theta_{ax}) = a'_x \exp(j\theta'_{ax}) = d\bar{v}_x / d\bar{t} = d^2 \bar{x} / d\bar{t}^2 \quad (84)$$

where

$$a'_x = -a_x \quad (85)$$

$$\theta'_{ax} = \theta_{ax} + \pi \quad (86)$$

so that an equivalent representation of the acceleration is

$$a'_x = -\sec \beta_{v xv x} \sin \beta_{tt}^x t^{-1} dv_x / d\theta_t^x \quad (87)$$

$$= -\sec \beta_{v xv x} \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \sec \beta_{xx} t^{-1} dx / d\theta_t^x) \quad (88)$$

$$\sim -\sec \beta_{v xv x} \sin^2 \beta_{tt}^x \sec \beta_{xx} t^{-2} d^2 x / d\theta_t^{x2} \quad (89)$$

$$\theta'_{ax} = \theta_x + \beta_{xx} + \beta_{v xv x} - 2(\theta_t^x + \beta_{tt}^x) + \pi \quad (90)$$

In the case of incoherent space and coherent time, equations (79), (88) and (90) give

$$a_x^{ic} = - \sec \beta_{vxvx}^{ic} t^{-2} d^2 x / d\theta_t^{x2} \quad (91)$$

$$\theta_{ax}^{ic} = \beta_{vxvx}^{ic} - 2\theta_t^x \quad (92)$$

where θ_{ax}^{ic} is a small number, and where from equations (37), (39), (51) and (79) it follows that

$$v_x^{ic} = t^{-1} dx / d\theta_t^x \quad (93)$$

$$\theta_{vx}^{ic} = - \theta_t^x - \pi/2 \quad (94)$$

$$\tan \beta_{vxvx}^{ic} = E_{xt}^{ic} / F_{xt}^{ic} \quad (95)$$

where

$$E_{xt}^{ic} = - dx / d\theta_t^x \quad E_{xt}^{ic} \geq 0 \quad (96)$$

$$F_{xt}^{ic} = d^2 x / d\theta_t^{x2} \quad F_{xt}^{ic} \geq 0 \quad (97)$$

$$\sec \beta_{vxvx}^{ic} = [(E_{xt}^{ic})^2 + (F_{xt}^{ic})^2]^{1/2} / F_{xt}^{ic} \quad (98)$$

Because $E_{xt}^{ic} \geq 0$ and $F_{xt}^{ic} \geq 0$ it follows that β_{vxvx}^{ic} is a small positive number for this case. Equations (91) and (98) give

$$a_x^{ic} = - t^{-2} [(E_{xt}^{ic})^2 + (F_{xt}^{ic})^2]^{1/2} \quad (99)$$

Case d. Coherent Space and Coherent Time.

In this section an equation for the acceleration of a particle is obtained which can be used to attain the limit of coherent space and coherent time which is defined by

$$\beta_{xx} = \pi/2 \quad \beta_{tt}^x = \pi/2 \quad (100)$$

The general expression for the magnitude and internal phase angle of the acceleration is obtained from equation (42) and (55) as

$$a_x = \csc \beta_{vxvx} \sin \beta_{tt}^x v_x / t d\theta_{vx} / d\theta_t^x \quad (101)$$

$$\theta_{ax} = \theta_x + \beta_{vxvx} + \beta_{xx} - 2(\theta_t^x + \beta_{tt}^x) \quad (102)$$

Combining equations (38) and (101) gives

$$a_x = \csc \beta_{vxvx} \sin^2 \beta_{tt}^x \csc \beta_{xx} x / t^2 d\theta_x / d\theta_t^x d\theta_{vx} / d\theta_t^x \quad (103)$$

where from equation (39) it follows that

$$d\theta_{vx}/d\theta_t^x = d\theta_x/d\theta_t^x - 1 + d/d\theta_t^x(\beta_{xx} - \beta_{tt}^x) \quad (104)$$

Another expression for the acceleration can be obtained from equations (38) and (80) which gives

$$a_x = \sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \csc \beta_{xx} x/t d\theta_x/d\theta_t^x) \quad (105)$$

In the present case it is convenient to write the acceleration in equation (42) as

$$\bar{a}_x = a_x \exp(j\theta_{ax}) = a'_x \exp(j\theta'_{ax}) = i\bar{v}_x/d\bar{t} = d^2\bar{x}/d\bar{t}^2 \quad (106)$$

where

$$a'_x = -a_x \quad (107)$$

$$\theta'_{ax} = \theta_{ax} + \pi \quad (108)$$

and an alternative representation of the acceleration is

$$a'_x = -\csc \beta_{vxvx} \sin^2 \beta_{tt}^x \csc \beta_{xx} x/t^2 d\theta_x/d\theta_t^x d\theta_{vx}/d\theta_t^x \quad (109)$$

$$= -\sec \beta_{vxvx} \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \csc \beta_{xx} x/t d\theta_x/d\theta_t^x) \quad (110)$$

$$\theta'_{ax} = \theta_{vx} + \beta_{vxvx} - \theta_t^x - \beta_{tt}^x + \pi \quad (111)$$

$$= \theta_x + \beta_{vxvx} + \beta_{xx} - 2(\theta_t^x + \beta_{tt}^x) + \pi$$

A comparison of equations (109) and (110) gives

$$\tan \beta_{vxvx} = C_{xt}/D_{xt} \quad (112)$$

$$\csc \beta_{vxvx} = (C_{xt}^2 + D_{xt}^2)^{1/2}/C_{xt} \quad (113)$$

$$\sec \beta_{vxvx} = (C_{xt}^2 + D_{xt}^2)^{1/2}/D_{xt} \quad (114)$$

where

$$C_{xt} = \sin \beta_{tt}^x \csc \beta_{xx} x/t d\theta_x/d\theta_t^x [d\theta_x/d\theta_t^x - 1 + d/d\theta_t^x(\beta_{xx} - \beta_{tt}^x)] \quad (115)$$

$$D_{xt} = d/d\theta_t^x (\sin \beta_{tt}^x \csc \beta_{xx} x/t d\theta_x/d\theta_t^x) \quad (116)$$

where in general $C_{xt} \leq 0$. Therefore

$$a'_x = - C_{xt}/t \csc \beta_{vxvx} \sin \beta_{tt}^x \quad (117)$$

$$= - D_{xt}/t \sec \beta_{vxvx} \sin \beta_{tt}^x \quad (118)$$

Equations (117) and (118) can also be written as

$$a'_x = - (C_{xt}^2 + D_{xt}^2)^{1/2}/t \sin \beta_{tt}^x \quad (119)$$

For the case of coherent space and coherent time equations (109) through (111) become with the help of equation (100)

$$a'_x = - \csc \beta_{vxvx}^c x/t^2 d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) \quad (120)$$

$$= - \sec \beta_{vxvx}^c x/t^2 d^2\theta_x/d\theta_t^{x2} \quad (121)$$

$$\theta'_{ax} = \theta_x + \beta_{vxvx}^c - 2\theta_t^x + \pi/2 \quad (122)$$

From equations (38) and (39) it follows that

$$v_x^c = x/t d\theta_x/d\theta_t^x \quad (123)$$

$$\theta_{vx}^c = \theta_x - \theta_t^x \quad (124)$$

Equations (51), (123) and (124) give

$$\tan \beta_{vxvx}^c = E_{xt}^c/F_{xt}^c \quad (125)$$

where

$$E_{xt}^c = d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) \quad E_{xt}^c \leq 0 \quad (126)$$

$$F_{xt}^c = d^2\theta_x/d\theta_t^{x2} \quad F_{xt}^c \geq 0 \quad (127)$$

$$\csc \beta_{vxvx}^c = [(E_{xt}^c)^2 + (F_{xt}^c)^2]^{1/2}/E_{xt}^c \quad (128)$$

$$\sec \beta_{vxvx}^c = [(E_{xt}^c)^2 + (F_{xt}^c)^2]^{1/2}/F_{xt}^c \quad (129)$$

where $d\theta_x/d\theta_t^x \leq 1$. In a gravitational field, for example, the following relationship holds¹

$$2\theta_t^x \sim 3\theta_x \quad d\theta_x/d\theta_t^x \sim 2/3 \quad (130)$$

Therefore in general $E_{xt}^c \leq 0$ and $F_{xt}^c \geq 0$, and it is convenient to introduce a new angle by writing

$$\beta_{vxvx}^c = -\pi/2 + \delta_{xt} \quad (131)$$

or equivalently

$$\tan \delta_{xt} = F_{xt}^c / |E_{xt}^c| \quad (132)$$

so that in general $\delta_{xt} \geq 0$. Combining equations (120) through (122) and (126) through (129) gives for coherent space and time

$$a_x^{c'} = -x/t^2 [(E_{xt}^c)^2 + (F_{xt}^c)^2]^{1/2} \quad (133)$$

$$\theta_{ax}^{c'} = \theta_{vx}^c + \beta_{vxvx}^c - \theta_t^x + \pi/2 \quad (134)$$

$$= \theta_{vx}^c - \theta_t^x + \delta_{xt}$$

$$= \theta_x + \beta_{vxvx}^c - 2\theta_t^x + \pi/2$$

$$= \theta_x - 2\theta_t^x + \delta_{xt}$$

Therefore $\theta_{ax}^{c'}$ is a small number. Actually equation (133) follows directly from equation (119) by noting that for coherent space and time equations (115) and (116) give

$$C_{xt}^c = x/t E_{xt}^c \quad D_{xt}^c = x/t F_{xt}^c \quad (135)$$

Coherent space and time represents an internal motion in space and time that can be written in complex number form as

$$d\bar{x} = j\bar{x}d\theta_x \quad d\bar{t} = j\bar{t}d\theta_t^x \quad (136)$$

which is equivalent to equation (100). The magnitude and phase angle equations (123) and (124) are equivalent to the following complex number expression for the particle speed in coherent spacetime

$$\bar{v}_x^c = (d\bar{x}/d\bar{t})^c = \bar{x}/\bar{t} d\theta_x/d\theta_t^x \quad (137)$$

The magnitude and phase angle of the acceleration that are given in equations (133) and (134) correspond to a complex number acceleration that is obtained from equation (106) and is given by

$$\bar{a}_x^c = (d^2\bar{x}/d\bar{t}^2)^c = \bar{x}/\bar{t}^2 [d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) - jd^2\theta_x/d\theta_t^{x2}] \quad (138)$$

as it must be because of the definition of δ_{xt} given in equations (131) and (132). For a linear solution of the form

$$\theta_x = a_x \theta_t^x + b_x \quad a_x \leq 1 \quad (139)$$

it follows from equation (133) that

$$a_x^{c'} = -x/t^2 |a_x(a_x - 1)| = -x/t^2 a_x(1 - a_x) \quad (140)$$

which corresponds to $F_{xt}^c = 0$ and $\beta_{v_x v_x}^c = -\pi/2$ by equations (125) through (127).

C. Harmonic Motion.

For harmonic motion

$$v_x = v_{x\omega} \exp(i\omega t) \quad (141)$$

and it follows from equation (23) that for harmonic motion in external space with x and t varying β_{xx} is an imaginary number given by

$$\beta_{xx} = -iQ_x \quad \tanh Q_x = \omega^{-1} \partial \theta_x / \partial t \quad (142)$$

$$\sec \beta_{xx} = \operatorname{sech} Q_x \quad \csc \beta_{xx} = i \operatorname{csch} Q_x \quad (143)$$

so that equations (35) through (38) become

$$v_x = i\omega x \cos \beta_{tt}^x \operatorname{sech} Q_x \quad (144)$$

$$= ix \cos \beta_{tt}^x \operatorname{csch} Q_x \, d\theta_x / dt \quad (145)$$

$$= i\omega x / t \sin \beta_{tt}^x \operatorname{sech} Q_x \, dt / d\theta_t^x \quad (146)$$

$$= ix / t \sin \beta_{tt}^x \operatorname{csch} Q_x \, d\theta_x / d\theta_t^x \quad (147)$$

where for harmonic motion the coordinates and velocities are complex numbers in external space so that

$$x = x_\omega \exp(i\omega t) \quad v_x = v_{x\omega} \exp(i\omega t) \quad v_{x\omega} = i\omega x_\omega \quad (148)$$

$$\bar{x} = x_\omega \exp(i\omega t + j\theta_x) \quad \bar{v}_x = v_{x\omega} \exp(i\omega t + j\theta_{v_x}) \quad (149)$$

Comparing equations (144) and (147) gives

$$\omega t = \tan \beta_{tt}^x \coth Q_x \, d\theta_x / d\theta_t^x \quad (150)$$

while combining equations (144) and (145) gives

$$\omega = \coth Q_x \, d\theta_x / dt \quad (151)$$

For the case of harmonic motion in external space, where x and t are

variables, β_{vxvx} is an imaginary number given by

$$\begin{aligned}\tan \beta_{vxvx} &= v_x \partial \theta_x / \partial v_x \\ &= x \partial \theta_{vx} / \partial x = -1/\omega \, d\theta_{vx} / dt\end{aligned}\quad (152)$$

Introducing a new quantity Q_{vx} by

$$\beta_{vxvx} = -iQ_{vx} \quad (153)$$

allows equation (152) to be written as

$$\tan \beta_{vxvx} = -i \tanh Q_{vx} = -1/\omega \, d\theta_{vx} / dt \quad (154)$$

$$\tanh Q_{vx} = \omega^{-1} \, d\theta_{vx} / dt \quad (155)$$

$$\sec \beta_{vxvx} = \operatorname{sech} Q_{vx} \quad \csc \beta_{vxvx} = 1 \operatorname{csch} Q_{vx} \quad (156)$$

which are valid for harmonic motion. For harmonic motion equation (53) can be rewritten using equation (143) as follows

$$a_x \sim -\omega^2 x \cos^2 \beta_{tt}^x \sec \beta_{vxvx} \operatorname{sech} Q_x \quad (157)$$

or since β_{vxvx} is an imaginary number given by equation (153) for the case where x and t are variables it follows from equation (156) that equation (157) can be written as

$$a_x \sim -\omega^2 x \cos^2 \beta_{tt}^x \operatorname{sech} Q_{vx} \operatorname{sech} Q_x \quad (158)$$

D. Measured and Conventionally Calculated Kinematic Quantities.

The average speed of a particle in spacetime with broken internal symmetries is defined as

$$\bar{v}_x^{av} = v_x^{av} \exp(j\theta_{vx}^{av}) = \bar{x}/\bar{t} = x/t \exp[j(\theta_x - \theta_t^x)] \quad (159)$$

$$v_x^{av} = x/t \quad \theta_{vx}^{av} = \theta_x - \theta_t^x \quad (160)$$

The conventional definition of the average speed is given by

$$v_{conx}^{av} = x_m/t_m^x = (x \cos \theta_x)/(t \cos \theta_t^x) \quad (161)$$

while the measured value of the average speed is obtained from equations (159) and (160)

$$v_{mx}^{av} = v_x^{av} \cos \theta_{vx}^{av} = x/t \cos(\theta_x - \theta_t^x) \quad (162)$$

A comparison of equations (161) and (162) gives

$$(v_{mx}^{av} - v_{conx}^{av})/v_x^{av} = \cos(\theta_x - \theta_t^x) - (\cos \theta_x)/(\cos \theta_t^x) \quad (163)$$

$$\sim \theta_t^x(\theta_x - \theta_t^x) \quad (164)$$

where equation (164) is valid for small angles.

The value of the instantaneous velocity is defined for broken symmetry spacetime by equations (34) through (39). The conventionally defined instantaneous velocity is given by

$$v_{conx} = dx_m/dt_m^x = d(x \cos \theta_x)/d(t \cos \theta_t^x) \quad (165)$$

$$= f_x dx/dt \cos \theta_x \sec \theta_t^x \quad (166)$$

where t_m^x is defined by equation (8), and where

$$f_x = (1 - \tan \theta_x \tan \beta_{xx})/(1 - \tan \theta_t^x \tan \beta_{tt}^x) \quad (167)$$

The measured value of the instantaneous speed of a particle in approximately incoherent spacetime is obtained from equations (34), (35), (39) and (40) to be

$$v_{mx} = v_x \cos \theta_{vx} \quad (168)$$

$$= dx/dt \cos \beta_{tt}^x \sec \beta_{xx} \cos \theta_{vx}$$

$$= v_{conx}/f_x \cos \beta_{tt}^x \sec \beta_{xx} \sec \theta_x \cos \theta_t^x \cos \theta_{vx}$$

where equation (35) was selected for v_x , and where $f_x \neq 0$. Combining equations (166) and (168) gives

$$(v_{mx} - v_{conx})/v_x = \cos(\theta_x + \beta_{xx} - \theta_t^x - \beta_{tt}^x) \quad (169)$$

$$- f_x \cos \theta_x \sec \theta_t^x \cos \beta_{xx} \sec \beta_{tt}^x$$

$$\sim (\theta_t^x + \beta_{tt}^x)(\theta_x + \beta_{xx} - \theta_t^x - \beta_{tt}^x)$$

For the case where $\beta_{tt}^x = 0$ and $\beta_{xx} = 0$ this equation becomes the same as equation (164) for the average speeds. Equations (165) and (166) show that a conventionally defined stationary particle is defined

$$v_{conx} = dx_m/dt_m = 0 \quad (170)$$

Generally this is associated with $dx/dt = 0$, but if it is associated with $f_x = 0$ then

$$\cot \theta_x = \tan \beta_{xx} \quad dx/x = \tan \theta_x d\theta_x \quad \beta_{xx} = \pi/2 - \theta_x \quad (171)$$

which integrates to

$$x = k \sec \theta_x \quad x_m = x \cos \theta_x = k \quad (172)$$

where $k = \text{constant}$. For this case $dx/dt \neq 0$ and in fact from equation (172)

$$dx/dt = k \sec \theta_x \tan \theta_x d\theta_x/dt \quad (173)$$

For the case $f_x = 0$, which corresponds to $v_{\text{con}x} = 0$ from equation (166) and which yields the relation in equation (173), it follows from equation (35) that the particle velocity magnitude and the measured particle velocity are given respectively by

$$v_x = k \cos \beta_{tt}^x \sec \beta_{xx} \sec \theta_x \tan \theta_x d\theta_x/d\theta_t^x \quad (174)$$

$$v_{mx} = k \cos \beta_{tt}^x \sec \beta_{xx} \sec \theta_x \tan \theta_x \cos \theta_{vx} d\theta_x/d\theta_t^x \quad (175)$$

where θ_{vx} is given by equation (39).

The conventionally defined instantaneous velocity given by equation (165) can also be written as

$$v_{\text{con}x} = g_x x/t d\theta_x/d\theta_t^x \sin \theta_x \csc \theta_t^x \quad (176)$$

where

$$g_x = (1 - \cot \theta_x x^{-1} dx/d\theta_x) / (1 - \cot \theta_t^x t^{-1} dt/d\theta_t^x) \quad (177)$$

In this case the measured instantaneous speed is obtained from equations (38) and (40) to be

$$v_{mx} = v_x \cos \theta_{vx} \quad (178)$$

$$= x/t d\theta_x/d\theta_t^x \csc \beta_{xx} \sin \beta_{tt}^x \cos \theta_{vx}$$

$$= v_{\text{con}x}/g_x \csc \beta_{xx} \sin \beta_{tt}^x \csc \theta_x \sin \theta_t^x \cos \theta_{vx}$$

where $g_x \neq 0$.

The particle momentum is written as

$$\bar{p}_x = p_x \exp(j\theta_{px}) = m\bar{v}_x = m d\bar{x}/d\bar{t} \quad (179)$$

$$p_x = mv_x = m \sec \beta_{xx} \cos \beta_{tt}^x dx/dt \quad (180)$$

$$\theta_{px} = \theta_{vx} = \theta_x + \beta_{xx} - \theta_t^x - \beta_{tt}^x \quad (181)$$

The measured momentum is

$$p_{mx} = p_x \cos \theta_{px} \quad (182)$$

The conventional momentum is

$$p_{conx} = mv_{conx} = m dx_m / dt_m \quad (183)$$

Then

$$p_{mx} = m_{px} v_{conx} \quad (184)$$

where

$$m_{px} = m/f_x \cos \beta_{tt}^x \sec \beta_{xx} \sec \theta_x \cos \theta_t^x \cos \theta_{px} \quad (185)$$

where $f_x \neq 0$, and where m_{px} = effective mass which relates the measured momentum to the conventional definition of particle speed in the x direction.

The average acceleration of a particle in broken symmetry spacetime is given by

$$\bar{a}_x^{av} = a_x^{av} \exp(j\theta_{ax}^{av}) = \bar{v}_x / \bar{t} = v_x / t \exp[j(\theta_{vx} - \theta_t^x)] \quad (186)$$

$$a_x^{av} = v_x / t \quad \theta_{ax}^{av} = \theta_{vx} - \theta_t^x \quad (187)$$

where \bar{v}_x = speed of particle after time \bar{t} for a particle starting from a rest position. The conventional definition of average acceleration for a particle initially at rest is given by

$$a_{conx}^{av} = v_{mx} / t_m = (v_x \cos \theta_{vx}) / (t \cos \theta_t^x) \quad (188)$$

The measured value of the average acceleration is obtained from equation (186) to be

$$a_{mx}^{av} = a_x^{av} \cos \theta_{ax}^{av} = v_x / t \cos(\theta_{vx} - \theta_t^x) \quad (189)$$

From equations (186) through (189) it follows that

$$(a_{mx}^{av} - a_{conx}^{av}) / a_x^{av} = \cos(\theta_{vx} - \theta_t^x) - (\cos \theta_{vx}) / (\cos \theta_t^x) \quad (190)$$

$$\sim \theta_t^x (\theta_{vx} - \theta_t^x) \quad (191)$$

where equation (191) is valid for small angles. Therefore in an external field which breaks the spacetime symmetry the measured average acceleration does not have the same value as the conventionally calculated average acceleration.

The instantaneous acceleration is given by equations (42) through (158). The conventionally defined acceleration is expressed in terms of measured space and time coordinates as follows

$$\begin{aligned} a_{\text{conx}} &= dv_{\text{mx}}/dt_m = d(v_x \cos \theta_{\text{vx}})/d(t \cos \theta_t^x) \\ &= h_x dv_x/dt \cos \theta_{\text{vx}} \sec \theta_t^x \end{aligned} \quad (192)$$

where

$$h_x = (1 - \tan \theta_{\text{vx}} \tan \beta_{\text{vxvx}})/(1 - \tan \theta_t^x \tan \beta_{\text{tt}}^x) \quad (193)$$

where from equation (35)

$$dv_x/dt = d/dt(\cos \beta_{\text{tt}}^x \sec \beta_{\text{xx}} dx/dt) \quad (194)$$

$$\sim \cos \beta_{\text{tt}}^x \sec \beta_{\text{xx}} d^2x/dt^2 \quad (195)$$

Combining equations (192) and (195) gives

$$a_{\text{conx}} \sim h_x \cos \beta_{\text{tt}}^x \sec \beta_{\text{xx}} \cos \theta_{\text{vx}} \sec \theta_t^x d^2x/dt^2 \quad (196)$$

On the other hand the measured value of the acceleration of a particle in broken symmetry spacetime is obtained from equations (49), (50), (53) and (55) to be

$$a_{\text{mx}} = a_x \cos \theta_{\text{ax}} \quad (197)$$

$$= \cos \beta_{\text{tt}}^x \sec \beta_{\text{vxvx}} \cos \theta_{\text{ax}} dv_x/dt \quad (198)$$

$$\sim \cos^2 \beta_{\text{tt}}^x \sec \beta_{\text{vxvx}} \sec \beta_{\text{xx}} \cos \theta_{\text{ax}} d^2x/dt^2 \quad (199)$$

where the internal phase angle of the acceleration is taken from equations (50) or (55). Clearly a_{conx} and a_{mx} are not equivalent. Combining equations (49), (50), (55), (192) and (198) gives

$$(a_{\text{mx}} - a_{\text{conx}})/a_x = \cos \theta_{\text{ax}} - h_x \cos \theta_{\text{vx}} \sec \theta_t^x \sec \beta_{\text{tt}}^x \cos \beta_{\text{vxvx}} \quad (200)$$

$$\sim (\theta_t^x + \beta_{\text{tt}}^x)(\theta_{\text{vx}} + \beta_{\text{vxvx}} - \theta_t^x - \beta_{\text{tt}}^x) \quad (201)$$

$$= (\theta_t^x + \beta_{\text{tt}}^x)[\theta_x + \beta_{\text{xx}} + \beta_{\text{vxvx}} - 2(\theta_t^x + \beta_{\text{tt}}^x)] \quad (202)$$

where equations (201) and (202) are valid for small internal phase angles. From equations (192) and (197) it follows that

$$a_{\text{mx}} = a_{\text{conx}}/h_x \cos \beta_{\text{tt}}^x \sec \beta_{\text{vxvx}} \sec \theta_{\text{vx}} \cos \theta_t^x \cos \theta_{\text{ax}} \quad (203)$$

$$= dv_x/dt \cos \beta_{\text{tt}}^x \sec \beta_{\text{vxvx}} \cos \theta_{\text{ax}} \quad (204)$$

where θ_{ax} is given by equations (50) or (55), and $h_x \neq 0$.

For the case when $h_x = 0$ and the conventional acceleration given by equation (192) is zero, $a_{conx} = 0$, then it follows from equation (193) that

$$\tan \beta_{vxvx} = v_x \partial \theta_{vx} / \partial v_x = \cot \theta_{vx} \quad \beta_{vxvx} = \pi/2 - \theta_{vx} \quad (205)$$

$$v_x = k \sec \theta_{vx} \quad v_{xm} = v_x \cos \theta_{vx} = k \quad (206)$$

Combining equations (49) and (206) gives

$$a_x = k \cos \beta_{tt}^x \sec \beta_{vxvx} \sec \theta_{vx} \tan \theta_{vx} d\theta_{vx}/dt \quad (207)$$

The measured acceleration is then given by

$$a_{mx} = a_x \cos \theta_{ax} \quad (208)$$

where θ_{ax} is given by equations (50) and (55). This is an example of a case where a force exerted on a body is associated with a zero value of acceleration in the conventional sense because $a_{conx} = 0$, but does have a measured acceleration value given by equation (208). The acceleration sensed by dynamical measurements is given by equation (208) and (47).

E. Measured and Conventional Angular Velocity.

The conventional angular speed is given by

$$\omega_{con} = d\phi_m/dt_m = f_\phi \cos \theta_\phi \sec \theta_t^\phi d\phi/dt \quad (209)$$

where

$$\phi_m = \phi \cos \theta_\phi \quad t_m = t \cos \theta_t^\phi \quad (210)$$

$$f_\phi = (1 - \tan \beta_{\phi\phi} \tan \theta_\phi) / (1 - \tan \beta_{tt}^\phi \tan \theta_t^\phi) \quad (211)$$

where θ_t^ϕ and β_{tt}^ϕ are the time internal phase angles associated with rotational motion. If $f_\phi = 0$ then $\omega_{con} = 0$ and

$$\tan \beta_{\phi\phi} = \cot \theta_\phi \quad \phi \partial \theta_\phi / \partial \phi = \cot \theta_\phi \quad \beta_{\phi\phi} = \pi/2 - \theta_\phi \quad (212)$$

or

$$\phi = k \sec \theta_\phi \quad \phi_m = k \quad (213)$$

Therefore a rotating system may be conventionally interpreted to be at rest, but rotates internally so that the magnitude of the angle changes in the following manner from equation (213)

$$d\phi/dt = k \sec \theta_\phi \tan \theta_\phi d\theta_\phi/dt \quad (214)$$

The complex number angular speed is written as

$$\bar{\omega} = \omega \exp(j\theta_\omega) = d\bar{\phi}/d\bar{t} \quad (215)$$

and the measured angular speed is given by equations (10) and (16) as

$$\begin{aligned} \omega_m &= \sec \beta_{\phi\phi} \cos \beta_{tt}^\phi \cos \theta_\omega d\phi/dt \\ &= k \cos \beta_{tt}^\phi \sec \beta_{\phi\phi} \cos \theta_\omega \sec \theta_\phi \tan \theta_\phi d\theta_\phi/dt \end{aligned} \quad (216)$$

where from equation (14)

$$\theta_\omega = \theta_\phi + \beta_{\phi\phi} - \theta_t - \beta_{tt}^\phi \quad (217)$$

where $\beta_{\phi\phi}$ and β_{tt}^ϕ are given by equations (4) and (5) respectively.

The angular momentum is written as

$$\bar{L} = L \exp(j\theta_L) = \bar{I} d\bar{\phi}/d\bar{t} \quad (218)$$

where the moment of inertia is written as

$$\bar{I} = I \exp(j\theta_I) \quad I = mr^2 \quad \theta_I = 2\theta_r \quad (219)$$

The measured value of the moment of inertia is given by

$$I_m = I \cos \theta_I \quad (220)$$

From equations (10), (14) and (218) it follows that the magnitude and internal phase angle of the angular momentum is given by

$$L = I \sec \beta_{\phi\phi} \cos \beta_{tt}^\phi d\phi/dt \quad (221)$$

$$\theta_L = \theta_I + \theta_\phi + \beta_{\phi\phi} - \theta_t - \beta_{tt}^\phi = \theta_I + \theta_\omega \quad (222)$$

where from equation (4)

$$\tan \beta_{\phi\phi} = \phi \partial \theta_\phi / \partial \phi \quad (223)$$

The measured angular momentum is given by

$$L_m = L \cos \theta_L \quad (224)$$

$$= I \sec \beta_{\phi\phi} \cos \beta_{tt}^\phi \cos \theta_L d\phi/dt \quad (225)$$

Combining equations (209) through (225) gives

$$L_m = I_{eff} d\phi_m/dt_m = I_{eff} \omega_{con} \quad (226)$$

where the effective moment of inertia is given by

$$I_{eff} = I_m / f_\phi \sec \beta_{\phi\phi} \cos \beta_{tt}^\phi \sec \theta_I \cos \theta_t^\phi \sec \theta_\phi \cos \theta_L \quad (227)$$

for $f_\phi \neq 0$. For the case $f_\phi = 0$ the conventional angular speed in equation (209) has a zero value, while equations (214), (221), (224) and (225) give

$$L = kI \sec \beta_{\phi\phi} \cos \beta_{tt}^\phi \sec \theta_\phi \tan \theta_\phi d\theta_\phi/dt \quad (228)$$

$$L_m = kI \sec \beta_{\phi\phi} \cos \beta_{tt}^\phi \sec \theta_\phi \tan \theta_\phi \cos \theta_L d\theta_\phi/dt \quad (229)$$

Equations (226) and (227) are not valid for the case $f_\phi = 0$.

3. NEWTON'S LAW OF MOTION IN BROKEN SYMMETRY SPACETIME. Newton's law of motion for a particle in a potential field can be written for spacetime with broken internal symmetries as follows⁴⁻⁶

$$m\bar{a}_x = m d^2\bar{x}/d\bar{t}^2 = -\partial\bar{W}/\partial\bar{x} = \bar{F}_x \quad (230)$$

where m = particle mass and \bar{W} = complex number potential which can be written as

$$\bar{W} = W \exp(j\theta_W) \quad (231)$$

The derivatives of the potential function can be written in four ways. Representing the complex number force in the following way

$$\bar{F}_x = F_x \exp(j\theta_{Fx}) = -\partial\bar{W}/\partial\bar{x} \quad (232)$$

gives

$$F_x = -\cos \beta_{xx} \sec \beta_{WW} \partial W/\partial x \quad (233)$$

$$= -\cos \beta_{xx} \csc \beta_{WW} W \partial\theta_W/\partial x \quad (234)$$

$$= -\sin \beta_{xx} \sec \beta_{WW} x^{-1} \partial W/\partial\theta_x \quad (235)$$

$$= -\sin \beta_{xx} \csc \beta_{WW} W/x \partial\theta_W/\partial\theta_x \quad (236)$$

and

$$\theta_{Fx} = \theta_W + \beta_{WW} - \theta_x - \beta_{xx} \quad (237)$$

where θ_{Fx} is a small angle, and where

$$\tan \beta_{WW} = W \partial \theta_W / \partial W \quad (238)$$

Then the phase angle condition for Newton's law is obtained for nearly incoherent space and nearly incoherent time from equations (55) and (237) as

$$\theta_{ax} = \theta_{Fx} \quad (239)$$

For nearly coherent space and nearly incoherent time, equations (63) and (237) give

$$\theta_{ax}^+ = \theta_{Fx} \quad (240)$$

For nearly incoherent space and nearly coherent time, equations (86) and (237) give

$$\theta_{ax}' = \theta_{Fx} \quad (241)$$

For the case of nearly coherent space and nearly coherent time it follows from equations (108) and (237) that

$$\theta_{ax}' = \theta_{Fx} \quad (242)$$

Newton's law of motion given by equation (230) will now be considered for the four kinematic spacetime conditions that were considered in Section 2.

A. Incoherent Space and Incoherent Time.

This section develops the Newtonian law of dynamics for broken symmetry spacetime in a form that is suitable to make the transition to the case of incoherent space and incoherent time which is described by

$$\theta_x = 0 \quad \beta_{xx} = 0 \quad \theta_t = 0 \quad \beta_{tt}^x = 0 \quad (243)$$

where x and t are variables. Combining equations (43), (49), (50), (230), (233) and (237) gives

$$m \cos \beta_{tt}^x \sec \beta_{vxvx} dv_x/dt = - \cos \beta_{xx} \sec \beta_{WW} \partial W / \partial x \quad (244)$$

$$\begin{aligned} \theta_{ax} &= \theta_{vx} + \beta_{vxvx} - \theta_t^x - \beta_{tt}^x \\ &= \theta_W + \beta_{WW} - \theta_x - \beta_{xx} \end{aligned} \quad (245)$$

where β_{vxvx} and β_{WW} are given by equations (51) and (238) respectively. Combining equations (35) and (244) gives

$$\begin{aligned} m \cos \beta_{tt}^x \sec \beta_{vxvx} d/dt (\cos \beta_{tt}^x \sec \beta_{xx} dx/dt) \\ = - \cos \beta_{xx} \sec \beta_{WW} \partial W / \partial x \end{aligned} \quad (246)$$

which can be written approximately as

$$m \cos^2 \beta_{tt}^x \sec \beta_{vxvx} \sec \beta_{xx} d^2x/dt^2 \sim - \cos \beta_{xx} \sec \beta_{ww} \partial W / \partial x \quad (247)$$

Combining equations (55) and (245) gives

$$\begin{aligned} \theta_{ax} &= \theta_x + \beta_{xx} + \beta_{vxvx} - 2(\theta_t^x + \beta_{tt}^x) \\ &= \theta_w + \beta_{ww} - \theta_x - \beta_{xx} \end{aligned} \quad (248)$$

If the potential function is given by a power law $W \sim x^{-\sigma}$ then equation (238) gives

$$\beta_{ww} = \beta_{xx} \quad (249)$$

Therefore a reasonable approximation to equation (248) for many potential functions is given by

$$2\theta_x + \beta_{xx} + \beta_{vxvx} - 2(\theta_t^x + \beta_{tt}^x) \sim \theta_w \quad (250)$$

With the further approximation that $\beta_{vxvx} \sim \beta_{xx}$ equation (250) becomes

$$2(\theta_x + \beta_{xx} - \theta_t^x - \beta_{tt}^x) \sim \theta_w \quad (251)$$

Within the approximations $\beta_{vxvx} \sim \beta_{xx}$ and $\beta_{ww} \sim \beta_{xx}$ equations (55) and (248) becomes

$$\begin{aligned} \theta_{ax} &\sim \theta_x + 2\beta_{xx} - 2(\theta_t^x + \beta_{tt}^x) \\ &\sim \theta_w - \theta_x \end{aligned} \quad (252)$$

while equation (247) with $\beta_{ww} \sim \beta_{xx}$ becomes

$$m \cos^2 \beta_{tt}^x \sec \beta_{vxvx} \sec \beta_{xx} d^2x/dt^2 \sim - \partial W / \partial x \quad (253)$$

and with the further approximation that $\beta_{vxvx} \sim \beta_{xx}$ equation (253) becomes

$$m \cos^2 \beta_{tt}^x \sec^2 \beta_{xx} d^2x/dt^2 \sim - \partial W / \partial x \quad (254)$$

If on the other hand it is assumed that $\beta_{vxvx} \sim 0$, or equivalently $\delta_{xt} = \pi/2$, then equations (250), (55) and (253) become

$$2\theta_x + \beta_{xx} - 2(\theta_t^x + \beta_{tt}^x) \sim \theta_w \quad (255)$$

$$\theta_{ax} \sim \theta_x + \beta_{xx} - 2(\theta_t^x + \beta_{tt}^x)$$

$$m \cos^2 \beta_{tt}^x \sec \beta_{xx} d^2x/dt^2 \sim - \partial W / \partial x \quad (257)$$

The approximations in equations (255) through (257) are not nearly as good as the approximations in equations (251), (252) and (254) respectively. The limiting case of incoherent space and incoherent time is obtained by setting all phase angles equal to zero and the exact equation (246) becomes

$$m d^2x/dt^2 = - \partial W/\partial x \quad (258)$$

which is the standard form of Newton's law of motion.

B. Coherent Space and Incoherent Time.

The case of coherent space and incoherent time is described by

$$\beta_{xx} = \pi/2 \quad \theta_t = 0 \quad \beta_{tt}^x = 0 \quad (259)$$

where θ_x and t are variables. This section develops a form of Newton's law of motion that is suitable for making the transition to the case of coherent space and incoherent time given by equation (259). Combining equations (64) and (236) gives

$$- m \csc \beta_{vxvx} \cos \beta_{tt}^x v_x d\theta_{vx}/dt = - \sin \beta_{xx} \csc \beta_{WW} W/x \partial \theta_W / \partial \theta_x \quad (260)$$

Equivalently, combining equations (65) and (236) gives

$$\begin{aligned} & - m \csc \beta_{vxvx} \cos^2 \beta_{tt}^x \csc \beta_{xx} x d\theta_x/dt d\theta_{vx}/dt \\ & = - \sin \beta_{xx} \csc \beta_{WW} W/x \partial \theta_W / \partial \theta_x \end{aligned} \quad (261)$$

where $d\theta_{vx}/dt$ is given by equation (60).

For the limiting case of coherent space and incoherent time given in equation (259), it follows that equation (261) becomes

$$- m \csc \beta_{vxvx}^{ci} x (d\theta_x/dt)^2 = - \csc \beta_{WW} W/x \partial \theta_W / \partial \theta_x \quad (262)$$

Using equations (71) through (74) with equation (262) gives

$$- mx[(E_{xt}^{ci})^2 + (F_{xt}^{ci})^2]^{1/2} = - \csc \beta_{WW} W/x \partial \theta_W / \partial \theta_x \quad (263)$$

where E_{xt}^{ci} and F_{xt}^{ci} are given by equations (72) and (73). If the potential function changes coherently when the space coordinate changes coherently it follows that $\beta_{WW} = \pi/2$ and equation (263) becomes

$$- mx[(E_{xt}^{ci})^2 + (F_{xt}^{ci})^2]^{1/2} = - W/x \partial \theta_W / \partial \theta_x \quad (264)$$

The internal phase angle equation for Newton's law of motion in coherent space and incoherent time is obtained from equations (78), (237) and (240) to be

$$\theta_{ax}^{ci+} = \theta_x + \kappa_{xt} = \theta_W - \theta_x \quad (265)$$

where κ_{xt} is given by equation (76).

C. Incoherent Space and Coherent Time.

In this section a form of Newton's dynamical law is developed for broken symmetry spacetime that can be used to attain the case of incoherent space and coherent time which is described by

$$\theta_x = 0 \quad \beta_{xx} = 0 \quad \beta_{tt}^x = \pi/2 \quad (266)$$

where x and θ_t^x are variables. Combining equations (87) and (233) gives

$$-m \sec \beta_{v xv x} \sin \beta_{tt}^x t^{-1} dv_x/d\theta_t^x = -\cos \beta_{xx} \sec \beta_{ww} \partial W/\partial x \quad (267)$$

From equations (88) and (233) it follows that equation (267) can be written as

$$\begin{aligned} & -m \sec \beta_{v xv x} \sin \beta_{tt}^x t^{-1} d/d\theta_t^x (\sin \beta_{tt}^x \sec \beta_{xx} t^{-1} dx/d\theta_t^x) \\ & = -\cos \beta_{xx} \sec \beta_{ww} \partial W/\partial x \end{aligned} \quad (268)$$

Equations (267) and (268) are completely general equations.

For the limiting case of incoherent space and coherent time, equation (268) becomes

$$-m \sec \beta_{v xv x}^{ic} t^{-2} d^2 x/d\theta_t^{x2} = -\sec \beta_{ww} \partial W/\partial x \quad (269)$$

Using equations (96) through (98) with equation (269) gives with $\beta_{ww} = 0$

$$-mt^{-2} [(E_{xt}^{ic})^2 + (F_{xt}^{ic})^2]^{1/2} = -\partial W/\partial x \quad (270)$$

where E_{xt}^{ic} and F_{xt}^{ic} are given by equations (96) and (97). The corresponding phase angle condition for Newton's law of motion is obtained from equations (92), (237) and (241) to be

$$\theta_{ax}^{ic'} = \beta_{v xv x}^{ic} - 2\theta_t^x = \theta_w \quad (271)$$

where $\theta_w = \text{constant}$ and $\beta_{ww} = 0$, and where $\beta_{v xv x}^{ic}$ is given by equation (95). Equations (270) and (271) represent Newton's dynamical law of motion for incoherent space and incoherent time.

D. Coherent Space and Coherent Time.

This section considers a formulation of Newton's law of dynamics in broken symmetry spacetime that can be reduced to the limiting case of totally coherent spacetime which is described by

$$\beta_{xx} = \pi/2 \quad \beta_{tt}^x = \pi/2 \quad (272)$$

where θ_x and θ_t are variables. The proper form of the law of motion for this case is obtained from equations (101), (106), (107), (230) and (236) which give

$$- m \csc \beta_{vxvx} \sin \beta_{tt}^x v_x/t d\theta_{vx}/d\theta_t^x \quad (273)$$

$$= - \csc \beta_{ww} \sin \beta_{xx} W/x \partial \theta_w / \partial \theta_x$$

Equation (273) can be rewritten using equation (109) as follows

$$- mx/t^2 \csc \beta_{vxvx} \sin^2 \beta_{tt}^x \csc \beta_{xx} d\theta_x/d\theta_t^x d\theta_{vx}/d\theta_t^x \quad (274)$$

$$= - \csc \beta_{ww} \sin \beta_{xx} W/x \partial \theta_w / \partial \theta_x$$

Equation (274) can be rewritten using equation (39) to give the following approximation for slowly varying β_{xx} and β_{tt}^x

$$- mx/t^2 \csc \beta_{vxvx} \sin^2 \beta_{tt}^x \csc \beta_{xx} d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) \quad (275)$$

$$\sim - \csc \beta_{ww} \sin \beta_{xx} W/x \partial \theta_w / \partial \theta_x$$

Combining equations (249) and (275) gives the further approximation

$$- mx/t^2 \csc \beta_{vxvx} \sin^2 \beta_{tt}^x \csc \beta_{xx} d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) \quad (276)$$

$$\sim - W/x \partial \theta_w / \partial \theta_x$$

Equations (110) and (236) give an equivalent form of Newton's law of motion

$$- mt^{-1} \sec \beta_{vxvx} \sin \beta_{tt}^x d/d\theta_t^x (\sin \beta_{tt}^x \csc \beta_{xx} x/t d\theta_x/d\theta_t^x) \quad (277)$$

$$= - \csc \beta_{ww} \sin \beta_{xx} W/x \partial \theta_w / \partial \theta_x$$

Equations (274) and (277) are equivalent to

$$- mt^{-1} (C_{xt}^2 + D_{xt}^2)^{1/2} \sin \beta_{tt}^x = - \csc \beta_{ww} \sin \beta_{xx} W/x \partial \theta_w / \partial \theta_x \quad (278)$$

$$\sim - W/x \partial \theta_w / \partial \theta_x$$

where C_{xt} and D_{xt} are given by equations (112) through (116). The corresponding phase angle equations (111), (237) and (242) give

$$\theta_x + \beta_{vxvx} + \beta_{xx} - 2(\theta_t + \beta_{tt}^x) + \pi = \theta_w + \beta_{ww} - \theta_x - \beta_{xx} \quad (279)$$

which is valid for nearly coherent space and nearly coherent time.

The case of coherent spacetime corresponds to equation (272), and equation (276) becomes

$$-mxt^{-2} \csc \beta_{v\bar{v}x}^c d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) = -W/x \partial\theta_W/\partial\theta_x \quad (280)$$

which can be rewritten using equations (125) through (128) as

$$-mxt^{-2} [(E_{xt}^c)^2 + (F_{xt}^c)^2]^{1/2} = -W/x \partial\theta_W/\partial\theta_x \quad (281)$$

where E_{xt}^c and F_{xt}^c are given by equations (126) and (127). Equation (281) is the coherent spacetime limit of equation (278). The phase angle condition for Newton's law of motion in coherent spacetime is obtained from equations (134), (237) and (242) or directly from equation (279) to be

$$\theta_{ax}^{c'} = \theta_x - 2\theta_t^x + \delta_{xt} = \theta_W - \theta_x \quad (282)$$

where δ_{xt} is given by equations (131) or (132). Equivalently, equation (282) can be rewritten as

$$2(\theta_x - \theta_t^x) + \delta_{xt} = \theta_W \quad (283)$$

Equations (281) through (283) are equivalent to the complex number form of Newton's dynamical law of motion given in equation (230) which for coherent spacetime is written as

$$m\bar{x}/\bar{t}^2 [d\theta_x/d\theta_t^x (d\theta_x/d\theta_t^x - 1) - jd^2\theta_x/d\theta_t^{x2}] = -\bar{W}/\bar{x} \partial\theta_W/\partial\theta_x \quad (284)$$

For a free particle moving in coherent spacetime

$$\partial\theta_W/\partial\theta_x = 0 \quad (285)$$

and two possible solutions to equation (284) can be found

$$\theta_x = c_1 \quad (286)$$

$$\theta_x = \theta_t^x + c_2 \quad (287)$$

where c_1 and c_2 are constants. These solutions can also be deduced from equation (281). Equation (286) represents a state of rest for internal motion, and equation (287) represents a state of uniform motion in internal spacetime.

E. Simple Harmonic Oscillator.

Perhaps the most elementary system in mechanics is the simple harmonic oscillator.⁴⁻⁶ For broken symmetry spacetime the complex number potential energy for the simple harmonic oscillator is given by

$$\bar{W} = 1/2\bar{k}\bar{x}^2 \quad W = 1/2kx^2 \quad \theta_W = \theta_k + 2\theta_x \quad (288)$$

The equations of motion for the simple harmonic oscillator are now considered for four limiting spacetime conditions.

Case a. Incoherent Space and Incoherent Time.

Equation (258) gives

$$d^2x/dt^2 + kx = 0 \quad (289)$$

which is the standard equation for the simple harmonic oscillator.

Case b. Coherent Space and Incoherent Time.

Equations (264) and (265) give

$$-m[(E_{xt}^{ci})^2 + (F_{xt}^{ci})^2]^{1/2} + k = 0 \quad (290)$$

$$\kappa_{xt} = \theta_k \quad (291)$$

where E_{xt}^{ci} and F_{xt}^{ci} are given by equations (72) and (73).

Case c. Incoherent Space and Coherent Time.

Equations (270) and (271) give

$$-mt^{-2}[(E_{xt}^{ic})^2 + (F_{xt}^{ic})^2]^{1/2} + kx = 0 \quad (292)$$

$$\beta_{v xv x}^{ic} - 2\theta_t^x = \theta_k \quad (293)$$

where E_{xt}^{ic} and F_{xt}^{ic} are given by equations (96) and (97).

Case d. Coherent Space and Coherent Time.

Equations (281) and (282) give

$$-mt^{-2}[(E_{xt}^c)^2 + (F_{xt}^c)^2]^{1/2} + k = 0 \quad (294)$$

$$\delta_{xt} - 2\theta_t^x = \theta_k \quad (295)$$

where E_{xt}^c and F_{xt}^c are given by equations (126) and (127). It is clear from these cases that the simple harmonic oscillator can undergo internal spacetime motion.

F. Measured Force.

The measured force is given by equations (43) through (47), (244) and (273) to be

$$F_{mx} = F_x \cos \theta_{Fx} = F_x \cos \theta_{ax} = m a_{mx} \quad (296)$$

$$= m \cos \beta_{tt}^x \sec \beta_{vxvx} \cos \theta_{ax} dv_x/dt \quad (297)$$

$$= m \sin \beta_{tt}^x \csc \beta_{vxvx} \cos \theta_{ax} v_x/t d\theta_{vx}/d\theta_t^x \quad (298)$$

Combining equations (203), (296) and (297) gives the measured force as

$$F_{mx} = m_{Fx} a_{conx} = m_{Fx} dv_{mx}/dt_m^x \quad (299)$$

with the effective mass m_{Fx} given by

$$m_{Fx} = m/h_x \cos \beta_{tt}^x \sec \beta_{vxvx} \sec \theta_{vx} \cos \theta_t^x \cos \theta_{ax} \quad (300)$$

where h_x is defined in equation (193). For broken symmetry spacetime the effective mass m_{Fx} relates the measured force to the conventionally defined acceleration in the x direction, and $m_{Fx} \neq m_{px}$ where m_{px} is given by equation (185). The measured time in the x direction is given by equation (8).

F. Conservation of Energy.

The obvious generalization of the law of conservation of energy to the case of broken symmetry spacetime is⁴⁻⁶

$$\bar{p}^2/(2m) + \bar{W} = \bar{E} \quad (301)$$

which can be rewritten as two scalar equations

$$p^2/(2m) \cos(2\theta_p) + W \cos \theta_W = E \cos \theta_E \quad (302)$$

$$p^2/(2m) \sin(2\theta_p) + W \sin \theta_W = E \sin \theta_E \quad (303)$$

or equivalently as

$$[p^2/(2m)]^2 + p^2/m W \cos(2\theta_p - \theta_W) + W^2 = E^2 \quad (304)$$

$$\tan(2\theta_p) = (E \sin \theta_E - W \sin \theta_W)/(E \cos \theta_E - W \cos \theta_W) \quad (305)$$

The value of the momentum magnitude obtained from equation (304) is

$$p^2 = -2mW \cos(2\theta_p - \theta_W) + 2m[E^2 - W^2 \sin^2(2\theta_p - \theta_W)]^{1/2} \quad (306)$$

In this way p and θ_p are obtained in terms of E , θ_E , W and θ_W . An approximate solution of equation (301) gives

$$p^2/(2m) + W \sim E \quad 2\theta_p \sim \theta_W \sim \theta_E \quad (307)$$

which assumes that each term in equation (301) has the same value of internal phase angle.

4. CONCLUSION. Broken spacetime symmetries are described by internal phase angles of the space and time coordinates, and affect the basic kinematic and dynamical equations of motion of Newtonian mechanics. The kinematic and dynamic variables such as particle speed, momentum, acceleration, force and energy must also have internal phase angles and be represented as complex numbers in an internal space. The internal phase angles of the space and time coordinates represent additional degrees of freedom for a particle and allow an internal motion to occur in matter for fixed magnitudes of the space and time coordinates. Elementary mechanical systems, such as the simple harmonic oscillator, can experience internal motions in space and time.

ACKNOWLEDGEMENT

I wish to thank Elizabeth K. Klein for her kind help in typing this paper.

REFERENCES

1. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.
2. Weiss, R. A., Relativistic Thermodynamics, Exposition Press, New York, 1976.
3. Weiss, R. A., "High- T_c Superconductivity and the Photoelectric Effect," Ninth Army Conference of Applied Mathematics and Computing, Univ. of Minnesota, Minneapolis, MN, ARO 92-1, June 18-21, 1991, p. 529.
4. Goldstein, H., Classical Mechanics, Addison-Wesley, New York, 1980.
5. Osgood, W. F., Mechanics, MacMillan, New York, 1949.
6. Corben, H. C. and Stehle, P., Classical Mechanics, John Wiley, New York, 1957.

SLOW AND ULTRAFAST WAVE PROPAGATION PROCESSES

Richard A. Weiss

U.S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. A theory of wave propagation in space and time with broken internal symmetries is developed from which the special cases of slow and ultrafast processes associated with the wave propagation can be obtained as limiting cases. Physical quantities and space and time coordinates are represented as complex numbers in an internal space. Ultrafast processes are associated with the coherent rotation of complex number physical quantities in an internal space while the magnitudes are held fixed. Slow processes are associated with changes in the magnitudes of the physical quantities. Spacetime can be coherent in which case the complex number space and time coordinates change by rotations in an internal space while the coordinate magnitudes remain fixed, or spacetime can be incoherent in which case the changes in spacetime coordinates occur as variations of the magnitudes of the coordinates. Eight possible special cases of wave propagation processes in asymmetric spacetime are delineated according to whether a process is slow or ultrafast, space is incoherent or coherent, and time is incoherent or coherent. The ultrafast processes described in this paper can be studied by femtosecond laser light pulses. The coherent spacetime condition is associated with the superconducting state of a high- T_c compound while the partially coherent spacetime condition is associated with the normal state of a high- T_c material.

1. INTRODUCTION. The intense interest in ultrafast processes has been stimulated by the development of very short (femtosecond) laser pulses as a diagnostic tool for studying processes that occur on short time scales.¹⁻¹² These processes include chemical reactions, optical dynamics of molecules, laser induced plasmas and their radiations, chemical explosions and many others.¹⁻¹³ Ultrafast laser sources have been developed that operate in the infrared, visible and ultraviolet regions of the electromagnetic spectrum. Incoherent femtosecond x ray emissions have been observed from laser induced plasmas, and can be used for studying ultrafast atomic and molecular processes in gases, liquids and solids.¹⁻⁸ The dynamical behavior of matter at femtosecond time scales gives a picture of molecular dynamics such as occurs in relaxation and transport phenomena.¹⁻¹² These studies can be made at surfaces and interfaces as well as within bulk matter. Specific phenomena that can be studied include: fluid transport, wave propagation and phonon interactions, diffusion and heat flow, electromagnetic emissions, vibrations, adsorption and desorption, phase transitions and chemical reactions such as detonations.¹⁻¹³ Processes that occur in supernova explosions such as the rapid neutron capture process, neutrino emissions and shock wave transmission also occur on short time scales.¹⁴⁻¹⁶ All of the above processes can occur in matter that is located in incoherent spacetime which appears in ordinary matter, or in coherent spacetime which occurs in the high- T_c superconducting state of matter or in matter located in very strong electromagnetic or gravitational fields. This paper studies the propagation of waves that are associated with short time scale energy transfers, and

develops the forms taken by the wave equation for slow and ultrafast wave propagation processes that occur in space and time that have broken internal symmetries due to a special structure of matter or to the presence of an external field such as gravitation or electromagnetism.

In the presence of external fields or in the vicinity of a peculiar atomic or molecular structure, the space and time coordinates must be represented as complex numbers as follows¹⁷

$$\bar{\alpha} = \alpha \exp(j\theta_{\alpha}) \quad (1)$$

where $\alpha = x, y, z$, and

$$\bar{t} = t \exp(j\theta_t) \quad (2)$$

All physical quantities, with the exception of the light speed in the vacuum, have broken internal symmetries and must be represented as complex numbers in an internal space.¹⁷ This includes, for example, pressure, entropy, energy and magnetic and electric field strengths. Therefore for the case of wave propagation the amplitude of the waves must be represented as a complex number in internal space as follows

$$\bar{\Psi} = \Psi \exp(j\theta_{\Psi}) \quad (3)$$

Strictly speaking, the value of the internal phase angle of the time is associated with the particular physical quantity which is varying with time, so that for the case at hand

$$\bar{t} = t \exp(j\theta_t^{\Psi}) \quad (4)$$

where θ_t^{Ψ} = internal phase angle of time that is associated with the time variation of the wave function Ψ . Space is taken to be homogeneous so that the internal phase angle of time θ_t^{Ψ} is independent of the internal phase angles θ_{α} of the space coordinates. In other cases, such as particle dynamics and kinematics, the space and time coordinates are not independent and the internal phase angles of the space coordinates θ_{α} are each associated with a corresponding internal phase angle of time θ_t^{α} for $\alpha = x, y, z$ with $\partial\theta_{\alpha}/\partial\theta_t^{\alpha} \neq 0$. But for the case of wave propagation the time and space coordinates are taken to be independent parameters so that θ_t^{Ψ} and θ_{α} are unrelated quantities with $\partial\theta_{\alpha}/\partial\theta_t^{\Psi} = 0$. In general

$$\Psi = \Psi(\alpha, \theta_{\alpha}, t, \theta_t^{\Psi}) \quad (5)$$

$$\theta_{\Psi} = \theta_{\Psi}(\alpha, \theta_{\alpha}, t, \theta_t^{\Psi}) \quad (6)$$

where $\alpha = x, y, z$. From equations (1), (2) and (3) it follows that

$$d\bar{\alpha} = \sec \beta_{\alpha\alpha} d\alpha \exp[j(\theta_{\alpha} + \beta_{\alpha\alpha})] \quad (7)$$

$$= \csc \beta_{\alpha\alpha} \alpha d\theta_{\alpha} \exp[j(\theta_{\alpha} + \beta_{\alpha\alpha})] \quad (8)$$

$$d\bar{t} = \sec \beta_{tt}^{\Psi} dt \exp[j(\theta_t^{\Psi} + \beta_{tt}^{\Psi})] \quad (9)$$

$$= \csc \beta_{tt}^{\Psi} t d\theta_t^{\Psi} \exp[j(\theta_t^{\Psi} + \beta_{tt}^{\Psi})] \quad (10)$$

$$d\bar{\Psi} = \sec \beta_{\Psi\Psi} d\Psi \exp[j(\theta_{\Psi} + \beta_{\Psi\Psi})] \quad (11)$$

$$= \csc \beta_{\Psi\Psi} \Psi d\theta_{\Psi} \exp[j(\theta_{\Psi} + \beta_{\Psi\Psi})] \quad (12)$$

where

$$\tan \beta_{\alpha\alpha} = \alpha \partial \theta_{\alpha} / \partial \alpha \quad (13)$$

$$\tan \beta_{tt}^{\Psi} = t \partial \theta_t^{\Psi} / \partial t \quad (14)$$

$$\tan \beta_{\Psi\Psi} = \Psi \partial \theta_{\Psi} / \partial \Psi \quad (15)$$

These expressions will be used in Sections 2 and 3 to obtain the first and second derivatives of the wave function with respect to the space and time coordinates.

This paper develops a general theory of wave propagation in space and time that have broken internal symmetries. The wave function must also be represented as a complex number in internal space, and therefore the possibility exists in nature of having slow wave propagation processes in which the wave function magnitude changes in space and time, and ultrafast wave propagation processes in which the wave function rotates in internal space with a constant magnitude. The space and time coordinates themselves can also change in an incoherent and a coherent manner, so that in fact there are eight possible limiting cases of wave propagation in asymmetric spacetime. The paper is briefly organized as follows: Section 2 evaluates the first and second derivatives of the wave function with respect to space and time coordinates; Section 3 develops the general forms of the wave equations that can be specialized to the eight limiting cases of slow and ultrafast processes, coherent and incoherent space, and coherent and incoherent time; and Section 4 gives the solutions to the eight limiting types of wave equations.

2. SPACE AND TIME DERIVATIVES. This section evaluates the space and time derivatives that enter the wave equation for space and time with broken internal symmetries. The wave equation for space and time with broken internal symmetries is written as the following generalization of the standard scalar wave equation¹⁸⁻²⁴

$$\sum_{\alpha} \partial^2 \bar{\Psi} / \partial \bar{\alpha}^2 = \bar{c}^{-2} \partial^2 \bar{\Psi} / \partial \bar{t}^2 \quad (16)$$

where the sum is over $\alpha = x, y, z$ and where the complex number space and time coordinates $\bar{\alpha}$ and \bar{t} are given by equations (1) and (2) respectively. The complex number wave function is written as

$$\bar{\Psi} = \Psi \exp(j\theta_{\Psi}) \quad (17)$$

and the complex number wave speed is written as

$$\bar{c} = c \exp(j\theta_c) \quad (18)$$

The complex number wave speed for sound waves can be written as¹⁷

$$\bar{c}^2 = \bar{K}/\rho = d\bar{P}/d\rho \quad (19)$$

where \bar{K} = complex number bulk modulus, and ρ = mass density. For electromagnetic waves in matter the wave speed is written as²⁵

$$\bar{c}^2 = (\bar{\epsilon}\bar{\mu})^{-1} \quad c^2 = (\epsilon\mu)^{-1} \quad 2\theta_c = -\theta_\epsilon - \theta_\mu \quad (20)$$

where $\bar{\epsilon}$ and $\bar{\mu}$ = complex number electric permittivity and magnetic permeability respectively. For electromagnetic waves in the vacuum with no external fields present $\theta_c = 0$.^{17,25} In general for wave propagation in matter θ_c is a small number. In order to delineate the various wave equations derived from equation (16) for the cases of slow and ultrafast processes and for various types of space and time variations, it is necessary to evaluate the second derivatives of the wave function with respect to space and time as are required by equation (16).

A. First Derivatives with Respect to Space and Time.

The first derivatives of the wave function with respect to space and time are written as

$$\bar{v}_\alpha = v_\alpha \exp(j\theta_{v\alpha}) = \partial\bar{\Psi}/\partial\bar{\alpha} \quad (21)$$

$$\bar{u} = u \exp(j\theta_u) = \partial\bar{\Psi}/\partial\bar{t} \quad (22)$$

where $\alpha = x, y, z$. For the space derivatives in equation (21) the magnitudes v_α can be written as

$$v_\alpha = \sec \beta_{\Psi\Psi} \cos \beta_{\alpha\alpha} \partial\Psi/\partial\alpha \quad (23)$$

$$= \csc \beta_{\Psi\Psi} \cos \beta_{\alpha\alpha} \Psi \partial\theta_\Psi/\partial\alpha \quad (24)$$

$$= \sec \beta_{\Psi\Psi} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial\Psi/\partial\theta_\alpha \quad (25)$$

$$= \csc \beta_{\Psi\Psi} \sin \beta_{\alpha\alpha} \Psi/\alpha \partial\theta_\Psi/\partial\theta_\alpha \quad (26)$$

and the internal phase angles as

$$\theta_{v\alpha} = \theta_\Psi + \beta_{\Psi\Psi} - \theta_\alpha - \beta_{\alpha\alpha} \quad (27)$$

where $\beta_{\alpha\alpha}$ and $\beta_{\Psi\Psi}$ are given by equations (13) and (15) respectively. The magnitude of the time derivative that appears in equation (22) is written as

$$u = \sec \beta_{\Psi\Psi} \cos \beta_{tt}^{\Psi} \partial\Psi/\partial t \quad (28)$$

$$= \csc \beta_{\Psi\Psi} \cos \beta_{tt}^{\Psi} \Psi \partial\theta_{\Psi}/\partial t \quad (29)$$

$$= \sec \beta_{\Psi\Psi} \sin \beta_{tt}^{\Psi} t^{-1} \partial\Psi/\partial\theta_t^{\Psi} \quad (30)$$

$$= \csc \beta_{\Psi\Psi} \sin \beta_{tt}^{\Psi} \Psi/t \partial\theta_{\Psi}/\partial\theta_t^{\Psi} \quad (31)$$

while the internal phase angle is given by

$$\theta_u = \theta_{\Psi} + \beta_{\Psi\Psi} - \theta_t^{\Psi} - \beta_{tt}^{\Psi} \quad (32)$$

where β_{tt}^{Ψ} is given by equation (14) and θ_t^{Ψ} = internal phase angle of time that is associated with the time variation of the wave amplitude Ψ .

Four limiting forms of the first derivative of the wave function with respect to the spatial coordinates will now be considered.

Case 1. Slow Process and Incoherent Space.

This is described by

$$\theta_{\Psi} = 0 \quad \beta_{\Psi\Psi} = 0 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad (33)$$

Then equations (23) and (27) give

$$v_{\alpha}^{si} = \partial\Psi/\partial\alpha \quad \theta_{v\alpha}^{si} = 0 \quad (34)$$

Case 2. Ultrafast Process and Incoherent Space.

This case is given by

$$\beta_{\Psi\Psi} = \pi/2 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad (35)$$

Equations (24) and (27) become for this case

$$v_{\alpha}^{ui} = \Psi \partial\theta_{\Psi}/\partial\alpha \quad \theta_{v\alpha}^{ui} = \theta_{\Psi} + \pi/2 \quad (36)$$

or equivalently

$$\bar{v}_{\alpha}^{ui} = j\bar{\Psi} \partial\theta_{\Psi}/\partial\alpha \quad (37)$$

Case 3. Slow Process and Coherent Space.

The following conditions hold for this case

$$\theta_{\Psi} = 0 \quad \beta_{\Psi\Psi} = 0 \quad \beta_{\alpha\alpha} = \pi/2 \quad (38)$$

and equations (25) and (27) become

$$v_{\alpha}^{sc} = \alpha^{-1} \partial \Psi / \partial \theta_{\alpha} \quad \theta_{v\alpha}^{sc} = -\theta_{\alpha} - \pi/2 \quad (39)$$

or equivalently

$$\bar{v}_{\alpha}^{sc} = -j/\bar{\alpha} \partial \Psi / \partial \theta_{\alpha} \quad (40)$$

Case 4. Ultrafast Process and Coherent Space.

This case is described by

$$\beta_{\Psi\Psi} = \pi/2 \quad \beta_{\alpha\alpha} = \pi/2 \quad (41)$$

and equations (26) and (27) give

$$v_{\alpha}^{uc} = \Psi/\alpha \partial \theta_{\Psi} / \partial \theta_{\alpha} \quad \theta_{v\alpha}^{uc} = \theta_{\Psi} - \theta_{\alpha} \quad (42)$$

or equivalently

$$\bar{v}_{\alpha}^{uc} = \bar{\Psi}/\bar{\alpha} \partial \theta_{\Psi} / \partial \theta_{\alpha} \quad (43)$$

These are the four limiting cases associated with the first derivative with respect to the spatial coordinates.

Now the four limiting conditions will be given for the first derivative of the wave function with respect to time.

Case 1. Slow Process and Incoherent Time.

This case is given by

$$\theta_{\Psi} = 0 \quad \beta_{\Psi\Psi} = 0 \quad \theta_t^{\Psi} = 0 \quad \beta_{tt}^{\Psi} = 0 \quad (44)$$

and equations (28) and (32) become

$$u^{si} = \partial \Psi / \partial t \quad \theta_u^{si} = 0 \quad (45)$$

Case 2. Ultrafast Process and Incoherent Time.

This case is described by

$$\beta_{\Psi\Psi} = \pi/2 \quad \theta_t^{\Psi} = 0 \quad \beta_{tt}^{\Psi} = 0 \quad (46)$$

Equations (29) and (32) then give

$$u^{ui} = \Psi \partial \theta_{\Psi} / \partial t \quad \theta_u^{ui} = \theta_{\Psi} + \pi/2 \quad (47)$$

or

$$\bar{u}^{ui} = j\bar{\Psi} \partial \theta_{\Psi} / \partial t \quad (48)$$

Case 3. Slow Process in Coherent Time.

The following conditions are valid for this case

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \beta_{tt}^{\psi} = \pi/2 \quad (49)$$

while equations (30) and (32) give

$$u^{sc} = t^{-1} \partial \psi / \partial \theta_t \quad \theta_u^{sc} = -\theta_t - \pi/2 \quad (50)$$

or

$$\bar{u}^{sc} = -j/\bar{t} \partial \psi / \partial \theta_t \quad (51)$$

Case 4. Ultrafast Process in Coherent Time.

This case is described by the following conditions

$$\beta_{\psi\psi} = \pi/2 \quad \beta_{tt}^{\psi} = \pi/2 \quad (52)$$

and equations (31) and (32) give

$$u^{uc} = \psi/t \partial \theta_{\psi} / \partial \theta_t^{\psi} \quad \theta_u^{uc} = \theta_{\psi} - \theta_t^{\psi} \quad (53)$$

which can be rewritten as

$$\bar{u}^{uc} = \bar{\psi}/\bar{t} \partial \theta_{\psi} / \partial \theta_t^{\psi} \quad (54)$$

B. Second Derivatives with Respect to Space.

The second derivatives of the wave function with respect to the spatial coordinates will now be represented in four general forms which can be specialized to four limiting cases of physical interest corresponding to slow and fast wave propagation processes in incoherent and coherent space. The second derivative of the wave function with respect to space coordinates is written as

$$\bar{\xi}_{\alpha} = \xi_{\alpha} \exp(j\theta_{\xi\alpha}) = \partial^2 \bar{\psi} / \partial \bar{\alpha}^2 = \partial \bar{v}_{\alpha} / \partial \bar{\alpha} \quad (55)$$

where $\alpha = x, y, z$, and where \bar{v}_{α} is defined in equation (21).

Case 1. Slow Process in Incoherent Space.

A general expression for the second spatial derivative of the wave function will be derived which can be used to pass to the limit of a slow process in incoherent space which is described by

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad (56)$$

for $\alpha = x, y$ and z . Equations (23) and (55) give

$$\xi_{\alpha} = \sec \beta_{v\alpha} \cos \beta_{\alpha\alpha} \partial v_{\alpha} / \partial \alpha \quad (57)$$

$$= \sec \beta_{v\alpha} \cos \beta_{\alpha\alpha} \partial / \partial \alpha (\sec \beta_{\psi\psi} \cos \beta_{\alpha\alpha} \partial \psi / \partial \alpha) \quad (58)$$

$$\sim \sec \beta_{v\alpha} \cos^2 \beta_{\alpha\alpha} \sec \beta_{\psi\psi} \partial^2 \psi / \partial \alpha^2 \quad (59)$$

while equations (27) and (55) give

$$\theta_{\xi\alpha} = \theta_{v\alpha} + \beta_{v\alpha} - \theta_{\alpha} - \beta_{\alpha\alpha} \quad (60)$$

$$= \theta_{\psi} + \beta_{\psi\psi} + \beta_{v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \quad (61)$$

where $\beta_{v\alpha}$ is given by

$$\tan \beta_{v\alpha} = v_{\alpha} \partial \theta_{v\alpha} / \partial v_{\alpha} \quad (62)$$

where v_{α} is given by equation (23) and $\theta_{v\alpha}$ by equation (27). For this case $\theta_{\xi\alpha}$ is a small number. In the limiting case of a slow process in incoherent space equation (56) is valid and equations (58) and (61) become

$$\xi_{\alpha}^{si} = \partial^2 \psi / \partial \alpha^2 \quad \theta_{\xi\alpha}^{si} = 0 \quad (63)$$

which is the conventional result.

Case 2. Ultrafast Process in Incoherent Space.

This section derives a general equation for the second derivatives of the wave function with respect to the space coordinates which can be utilized to obtain the limiting case of an ultrafast process in incoherent space which is defined by

$$\beta_{\psi\psi} = \pi/2 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad (64)$$

where θ_{ψ} is now a variable. From equations (24) and (55) it follows that

$$\xi_{\alpha} = \csc \beta_{v\alpha} \cos \beta_{\alpha\alpha} v_{\alpha} \partial \theta_{v\alpha} / \partial \alpha \quad (65)$$

$$= \csc \beta_{v\alpha} \cos^2 \beta_{\alpha\alpha} \csc \beta_{\psi\psi} \psi \partial \theta_{\psi} / \partial \alpha \partial \theta_{v\alpha} / \partial \alpha \quad (66)$$

where $\beta_{v\alpha}$ is given by equation (62) with v_{α} given by equation (24) and where equation (27) gives

$$\partial \theta_{v\alpha} / \partial \alpha = \partial / \partial \alpha (\theta_{\psi} + \beta_{\psi\psi} - \theta_{\alpha} - \beta_{\alpha\alpha}) \quad (67)$$

The corresponding internal phase angle for the second spatial derivative is given by

$$\theta_{\xi\alpha} = \theta_{v\alpha} + \beta_{v\alpha v\alpha} - \theta_{\alpha} - \beta_{\alpha\alpha} \quad (68)$$

$$= \theta_{\psi} + \beta_{\psi\psi} + \beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \quad (69)$$

For an ultrafast process it is convenient to introduce an alternative representation of the second derivatives which is given by

$$\bar{\xi}_{\alpha} = \xi_{\alpha} \exp(j\theta_{\xi\alpha}) = \xi_{\alpha}^{\dagger} \exp(j\theta_{\xi\alpha}^{\dagger}) \quad (70)$$

$$= \partial^2 \bar{\Psi} / \partial \bar{\alpha}^2 = \partial \bar{v}_{\alpha} / \partial \bar{\alpha}$$

where

$$\xi_{\alpha}^{\dagger} = -\xi_{\alpha} \quad (71)$$

$$\theta_{\xi\alpha}^{\dagger} = \theta_{\xi\alpha} - \pi \quad (72)$$

Then it follows that

$$\xi_{\alpha}^{\dagger} = -\csc \beta_{v\alpha v\alpha} \cos \beta_{\alpha\alpha} v_{\alpha} \partial \theta_{v\alpha} / \partial \alpha \quad (73)$$

$$= -\csc \beta_{v\alpha v\alpha} \cos^2 \beta_{\alpha\alpha} \csc \beta_{\psi\psi} \psi \partial \theta_{\psi} / \partial \alpha \partial \theta_{v\alpha} / \partial \alpha \quad (74)$$

and

$$\theta_{\xi\alpha}^{\dagger} = \theta_{\psi} + \beta_{\psi\psi} + \beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) - \pi \quad (75)$$

for the general case.

In the limiting case of an ultrafast process in incoherent space, equation (64) is valid and equations (61) and (66) become

$$\xi_{\alpha}^{ui} = \csc \beta_{v\alpha v\alpha}^{ui} \psi (\partial \theta_{\psi} / \partial \alpha)^2 \quad (76)$$

$$\theta_{\xi\alpha}^{ui} = \theta_{\psi} + \beta_{v\alpha v\alpha}^{ui} + \pi/2 \quad (77)$$

From equations (36) and (62) it follows that

$$\tan \beta_{v\alpha v\alpha}^{ui} = E_{\psi\alpha}^{ui} / F_{\psi\alpha}^{ui} \quad (78)$$

where

$$E_{\psi\alpha}^{ui} = (\partial \theta_{\psi} / \partial \alpha)^2 \quad E_{\psi\alpha}^{ui} \geq 0 \quad (79)$$

$$F_{\psi\alpha}^{ui} = \partial^2 \theta_{\psi} / \partial \alpha^2 \quad F_{\psi\alpha}^{ui} \leq 0 \quad (80)$$

From equation (78) it follows that

$$\csc \beta_{v\alpha}^{ui} = [(E_{\psi\alpha}^{ui})^2 + (F_{\psi\alpha}^{ui})^2]^{1/2} / E_{\psi\alpha}^{ui} \quad (81)$$

Equations (76) and (81) give

$$\xi_{\alpha}^{ui} = \psi [(E_{\psi\alpha}^{ui})^2 + (F_{\psi\alpha}^{ui})^2]^{1/2} \quad (82)$$

For the signs chosen in equations (79) and (80) it follows from equation (78) that

$$\beta_{v\alpha}^{ui} = \pi/2 + \kappa_{\psi\alpha} \quad (83)$$

where $\kappa_{\psi\alpha}$ is a small positive number defined by

$$\tan \kappa_{\psi\alpha} = |F_{\psi\alpha}^{ui}| / E_{\psi\alpha}^{ui} \quad (84)$$

Equations (77) and (83) give

$$\theta_{\xi\alpha}^{ui} = \theta_{\psi} + \kappa_{\psi\alpha} + \pi \quad (85)$$

Finally equations (71) and (72) give

$$\xi_{\alpha}^{ui+} = -\psi [(E_{\psi\alpha}^{ui})^2 + (F_{\psi\alpha}^{ui})^2]^{1/2} \quad (86)$$

$$\theta_{\xi\alpha}^{ui+} = \theta_{\psi} + \kappa_{\psi\alpha} \quad (87)$$

so that $\theta_{\xi\alpha}^{ui+}$ is a small number.

Case 3. Slow Process in Coherent Space.

An expression is derived for the second derivative of the wave function with respect to the spatial coordinates, which can be used to pass to the limit of a slow process in coherent space whose characteristics are

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \beta_{\alpha\alpha} = \pi/2 \quad (88)$$

where θ_{α} is variable. From equation (55), (25) and (61) it follows that

$$\xi_{\alpha} = \sec \beta_{v\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial v_{\alpha} / \partial \theta_{\alpha} \quad (89)$$

$$= \sec \beta_{v\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial / \partial \theta_{\alpha} (\sec \beta_{\psi\psi} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial \psi / \partial \theta_{\alpha}) \quad (90)$$

$$\theta_{\xi\alpha} = \theta_{\psi} + \beta_{\psi\psi} + \beta_{v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \quad (91)$$

where $\beta_{v\alpha}$ is given by (25), (27) and (62). In this case it is convenient to introduce another representation of the second spatial derivative, namely

$$\begin{aligned}\bar{\xi}_\alpha &= \xi_\alpha \exp(j\theta_{\xi\alpha}) = \xi'_\alpha \exp(j\theta'_{\xi\alpha}) \\ &= \partial^2 \bar{\psi} / \partial \bar{\alpha}^2 = \partial \bar{v}_\alpha / \partial \bar{\alpha}\end{aligned}\quad (92)$$

where

$$\xi'_\alpha = -\xi_\alpha \quad (93)$$

$$\theta'_{\xi\alpha} = \theta_{\xi\alpha} + \pi \quad (94)$$

The limiting case of a slow process in coherent space is obtained from equation (88) which combined with equations (90) and (91) gives

$$\xi_\alpha^{sc} = \sec \beta_{v\alpha v\alpha}^{sc} \alpha^{-2} \partial^2 \psi / \partial \theta_\alpha^2 \quad (95)$$

$$\theta_{\xi\alpha}^{sc} = \beta_{v\alpha v\alpha}^{sc} - 2\theta_\alpha - \pi \quad (96)$$

Equations (39) and (62) give

$$\tan \beta_{v\alpha v\alpha}^{sc} = E_{\psi\alpha}^{sc} / F_{\psi\alpha}^{sc} \quad (97)$$

where

$$E_{\psi\alpha}^{sc} = -\partial \psi / \partial \theta_\alpha \quad E_{\psi\alpha}^{sc} \geq 0 \quad (98)$$

$$F_{\psi\alpha}^{sc} = \partial^2 \psi / \partial \theta_\alpha^2 \quad F_{\psi\alpha}^{sc} \geq 0 \quad (99)$$

$$\sec \beta_{v\alpha v\alpha}^{sc} = [(E_{\psi\alpha}^{sc})^2 + (F_{\psi\alpha}^{sc})^2]^{1/2} / F_{\psi\alpha}^{sc} \quad (100)$$

and therefore $\beta_{v\alpha v\alpha}^{sc}$ is a small positive angle. Equations (95), (98) and (100) give

$$\xi_\alpha^{sc} = \alpha^{-2} [(E_{\psi\alpha}^{sc})^2 + (F_{\psi\alpha}^{sc})^2]^{1/2} \quad (101)$$

The alternative description of the second derivative with respect to space is obtained from equations (93), (94), (96) and (101) as

$$\xi_\alpha^{sc'} = -\alpha^{-2} [(E_{\psi\alpha}^{sc})^2 + (F_{\psi\alpha}^{sc})^2]^{1/2} \quad (102)$$

$$\theta_{\xi\alpha}^{sc'} = \beta_{v\alpha v\alpha}^{sc} - 2\theta_\alpha \quad (103)$$

where $\theta_{\xi\alpha}^{sc'}$ is seen to be a small angle.

Case 4. Ultrafast Process and Coherent Space.

This section develops a representation for the second derivative of the wave function with respect to the space coordinates which can be utilized to

attain the limit of an ultrafast process in coherent space whose description is

$$\beta_{\psi\psi} = \pi/2 \quad \beta_{\alpha\alpha} = \pi/2 \quad (104)$$

The magnitude and internal phase angle of the second derivative is obtained from equations (26), (55), (61) to be

$$\xi_{\alpha} = \csc \beta_{v\alpha v\alpha} \sin \beta_{\alpha\alpha} v_{\alpha}/\alpha \partial \theta_{v\alpha}/\partial \theta_{\alpha} \quad (105)$$

$$= \csc \beta_{v\alpha v\alpha} \sin^2 \beta_{\alpha\alpha} \csc \beta_{\psi\psi} \psi/\alpha^2 \partial \theta_{\psi}/\partial \theta_{\alpha} \partial \theta_{v\alpha}/\partial \theta_{\alpha} \quad (106)$$

$$\theta_{\xi\alpha} = \theta_{\psi} + \beta_{\psi\psi} + \beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \quad (107)$$

where $\beta_{v\alpha v\alpha}$ is given by equation (26) and (62), and where from equation (27) it follows that

$$\partial \theta_{v\alpha}/\partial \theta_{\alpha} = \partial \theta_{\psi}/\partial \theta_{\alpha} - 1 + \partial/\partial \theta_{\alpha} (\beta_{\psi\psi} - \beta_{\alpha\alpha}) \quad (108)$$

A different expression for the second derivative can be obtained from equations (89) and (26) and is

$$\xi_{\alpha} = \sec \beta_{v\alpha v\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial v_{\alpha}/\partial \theta_{\alpha} \quad (109)$$

$$= \sec \beta_{v\alpha v\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial/\partial \theta_{\alpha} (\csc \beta_{\psi\psi} \sin \beta_{\alpha\alpha} \psi/\alpha \partial \theta_{\psi}/\partial \theta_{\alpha}) \quad (110)$$

A comparison of equations (106) and (110) gives

$$\tan \beta_{v\alpha v\alpha} = C_{\psi\alpha}/D_{\psi\alpha} \quad (111)$$

$$\csc \beta_{v\alpha v\alpha} = (C_{\psi\alpha}^2 + D_{\psi\alpha}^2)^{1/2}/C_{\psi\alpha} \quad (112)$$

$$\sec \beta_{v\alpha v\alpha} = (C_{\psi\alpha}^2 + D_{\psi\alpha}^2)^{1/2}/D_{\psi\alpha} \quad (113)$$

where

$$C_{\psi\alpha} = \sin \beta_{\alpha\alpha} \csc \beta_{\psi\psi} \psi/\alpha \partial \theta_{\psi}/\partial \theta_{\alpha} \partial \theta_{v\alpha}/\partial \theta_{\alpha} \quad (114)$$

$$D_{\psi\alpha} = \partial/\partial \theta_{\alpha} (\csc \beta_{\psi\psi} \sin \beta_{\alpha\alpha} \psi/\alpha \partial \theta_{\psi}/\partial \theta_{\alpha}) \quad (115)$$

where $\partial \theta_{v\alpha}/\partial \theta_{\alpha}$ is given by equation (108) and where $C_{\psi\alpha} \leq 0$. Therefore,

$$\xi_{\alpha} = C_{\psi\alpha} \alpha^{-1} \csc \beta_{v\alpha v\alpha} \sin \beta_{\alpha\alpha} \quad (116)$$

$$= D_{\psi\alpha} \alpha^{-1} \sec \beta_{v\alpha v\alpha} \sin \beta_{\alpha\alpha} \quad (117)$$

which can be rewritten as

$$\xi_{\alpha} = (C_{\Psi\alpha}^2 + D_{\Psi\alpha}^2)^{1/2} \alpha^{-1} \sin \beta_{\alpha\alpha} \quad (118)$$

It is convenient for the case at hand to write the second derivative of the wave function with respect to space coordinates in equation (55) as

$$\begin{aligned} \bar{\xi}_{\alpha} &= \xi_{\alpha} \exp(j\theta_{\xi\alpha}) = \xi'_{\alpha} \exp(j\theta'_{\xi\alpha}) \\ &= \partial^2 \bar{\Psi} / \partial \bar{\alpha}^2 = \partial \bar{V}_{\alpha} / \partial \bar{\alpha} \end{aligned} \quad (119)$$

where

$$\xi'_{\alpha} = - \xi_{\alpha} \quad (120)$$

$$\theta'_{\xi\alpha} = \theta_{\xi\alpha} + \pi \quad (121)$$

which gives a useful alternative description of the second derivative with respect to space.

For an ultrafast process in coherent space described by equation (104) it follows from equations (106), (110), (118) and (120) that

$$\xi_{\alpha}^{uc'} = - \csc \beta_{v\alpha v\alpha}^{uc} \Psi / \alpha^2 E_{\Psi\alpha}^{uc} \quad (122)$$

$$= - \sec \beta_{v\alpha v\alpha}^{uc} \Psi / \alpha^2 F_{\Psi\alpha}^{uc} \quad (123)$$

$$= - \alpha^{-1} [(C_{\Psi\alpha}^{uc})^2 + (D_{\Psi\alpha}^{uc})^2]^{1/2} \quad (124)$$

$$= - \Psi \alpha^{-2} [(E_{\Psi\alpha}^{uc})^2 + (F_{\Psi\alpha}^{uc})^2]^{1/2} \quad (125)$$

where

$$E_{\Psi\alpha}^{uc} = \partial \theta_{\Psi} / \partial \theta_{\alpha} (\partial \theta_{\Psi} / \partial \theta_{\alpha} - 1) \quad E_{\Psi\alpha}^{uc} \leq 0 \quad (126)$$

$$F_{\Psi\alpha}^{uc} = \partial^2 \theta_{\Psi} / \partial \theta_{\alpha}^2 \quad F_{\Psi\alpha}^{uc} \geq 0 \quad (127)$$

$$C_{\Psi\alpha}^{uc} = \Psi / \alpha E_{\Psi\alpha}^{uc} \quad (128)$$

$$D_{\Psi\alpha}^{uc} = \Psi / \alpha F_{\Psi\alpha}^{uc} \quad (129)$$

From equations (42) and (62) or directly from equation (111) it follows that

$$\tan \beta_{\psi\alpha}^{uc} = E_{\psi\alpha}^{uc} / F_{\psi\alpha}^{uc} \quad (130)$$

$$\csc \beta_{\psi\alpha}^{uc} = [(E_{\psi\alpha}^{uc})^2 + (F_{\psi\alpha}^{uc})^2]^{1/2} / E_{\psi\alpha}^{uc} \quad (131)$$

$$\sec \beta_{\psi\alpha}^{uc} = [(E_{\psi\alpha}^{uc})^2 + (F_{\psi\alpha}^{uc})^2]^{1/2} / F_{\psi\alpha}^{uc} \quad (132)$$

Because of the choice of signs in equations (126) and (127) it follows that

$$\beta_{\psi\alpha}^{uc} = -\pi/2 + \delta_{\psi\alpha} \quad (133)$$

where $\delta_{\psi\alpha} \geq 0$, so that

$$\tan \delta_{\psi\alpha} = F_{\psi\alpha}^{uc} / |E_{\psi\alpha}^{uc}| \quad (134)$$

The internal phase angle of the second derivative then follows from equations (104), (107), (121) and (133) as

$$\theta_{\xi\alpha}^{uc'} = \theta_{\psi} + \beta_{\psi\alpha}^{uc} - 2\theta_{\alpha} + \pi/2 \quad (135)$$

$$= \theta_{\psi} - 2\theta_{\alpha} + \delta_{\psi\alpha} \quad (136)$$

so that $\theta_{\xi\alpha}^{uc'}$ is a small number. The complex number second derivative with respect to space coordinates that corresponds to equations (125) and (136) is given by

$$\bar{\xi}_{\alpha}^{uc} = (\partial^2 \bar{\psi} / \partial \bar{\alpha}^2)^{uc} = \bar{\psi} / \bar{\alpha}^2 [\partial \theta_{\psi} / \partial \theta_{\alpha} (\partial \theta_{\psi} / \partial \theta_{\alpha} - 1) - j \partial^2 \theta_{\psi} / \partial \theta_{\alpha}^2] \quad (137)$$

for an ultrafast process in coherent spacetime.

C. Second Derivative with Respect to Time.

This section evaluates the second derivative of the wave function with respect to time for four cases of physical interest: slow and ultrafast processes and coherent and incoherent time. The complex number second derivative of the wave function with respect to time is written as

$$\bar{\xi}_t = \xi_t \exp(j\theta_{\xi t}) = \partial^2 \bar{\psi} / \partial \bar{t}^2 = \partial \bar{u} / \partial \bar{t} \quad (138)$$

where \bar{u} is defined in equation (22).

Case 1. Slow Process in Incoherent Time.

For this case a general expression for the second derivative of the wave function with respect to time is derived which allows a transition to the limiting case of a slow process in incoherent time which is described by

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \theta_t = 0 \quad \beta_{tt}^{\psi} = 0 \quad (139)$$

Equations (28), (32) and (138) give

$$\xi_t = \sec \beta_{uu} \cos \beta_{tt}^{\Psi} \partial u / \partial t \quad (140)$$

$$= \sec \beta_{uu} \cos \beta_{tt}^{\Psi} \partial / \partial t (\sec \beta_{\Psi\Psi} \cos \beta_{tt}^{\Psi} \partial \Psi / \partial t) \quad (141)$$

$$\sim \sec \beta_{uu} \cos^2 \beta_{tt}^{\Psi} \sec \beta_{\Psi\Psi} \partial^2 \Psi / \partial t^2 \quad (142)$$

$$\theta_{\xi t} = \theta_u + \beta_{uu} - \theta_t^{\Psi} - \beta_{tt}^{\Psi} \quad (143)$$

$$= \theta_{\Psi} + \beta_{\Psi\Psi} + \beta_{uu} - 2(\theta_t^{\Psi} + \beta_{tt}^{\Psi}) \quad (144)$$

where

$$\tan \beta_{uu} = u \partial \theta_u / \partial u \quad (145)$$

where u and θ_u are given by equations (28) and (32) respectively.

The limiting case of a slow process in incoherent time is obtained by imposing the conditions in equation (139) on equations (141) and (144) with the result that

$$\xi_t^{si} = \partial^2 \Psi / \partial t^2 \quad \theta_{\xi t}^{si} = 0 \quad (146)$$

which is the standard result.

Case 2. Ultrafast Process and Incoherent Time.

The general expression for the second derivative of the wave function with respect to time will now be given which produces the correct limit for the case of an ultrafast process and incoherent time which is described by

$$\beta_{\Psi\Psi} = \pi/2 \quad \theta_t = 0 \quad \beta_{tt}^{\Psi} = 0 \quad (147)$$

From equations (29), (32) and (138) it follows that

$$\xi_t = \csc \beta_{uu} \cos \beta_{tt}^{\Psi} u \partial \theta_u / \partial t \quad (148)$$

$$= \csc \beta_{uu} \cos^2 \beta_{tt}^{\Psi} \csc \beta_{\Psi\Psi} \Psi \partial \theta_{\Psi} / \partial t \partial \theta_u / \partial t \quad (149)$$

$$\theta_{\xi t} = \theta_u + \beta_{uu} - \theta_t^{\Psi} - \beta_{tt}^{\Psi} \quad (150)$$

$$= \theta_{\Psi} + \beta_{\Psi\Psi} + \beta_{uu} - 2(\theta_t^{\Psi} + \beta_{tt}^{\Psi}) \quad (151)$$

where β_{uu} is given by equation (145) with u and θ_u given by equations (29) and

(32) respectively, and where equation (32) gives

$$\partial \theta_u / \partial t = \partial / \partial t (\theta_\psi + \beta_{\psi\psi} - \theta_t^\psi - \beta_{tt}^\psi) \quad (152)$$

Define the following alternative representation of the second time derivative of the wave function as follows

$$\begin{aligned} \bar{\xi}_t &= \xi_t \exp(j\theta_{\xi t}) = \xi_t^\dagger \exp(j\theta_{\xi t}^\dagger) \\ &= \partial^2 \bar{\psi} / \partial \bar{t}^2 = \partial \bar{u} / \partial \bar{t} \end{aligned} \quad (153)$$

where

$$\xi_t^\dagger = -\xi_t \quad (154)$$

$$\theta_{\xi t}^\dagger = \theta_{\xi t} - \pi \quad (155)$$

so that

$$\xi_t^\dagger = -\csc \beta_{uu} \cos \beta_{tt}^\psi u \partial \theta_u / \partial t \quad (156)$$

$$= -\csc \beta_{uu} \cos^2 \beta_{tt}^\psi \csc \beta_{\psi\psi} \psi \partial \theta_\psi / \partial t \partial \theta_u / \partial t \quad (157)$$

$$\theta_{\xi t}^\dagger = \theta_\psi + \beta_{\psi\psi} + \beta_{uu} - 2(\theta_t^\psi + \beta_{tt}^\psi) - \pi \quad (158)$$

For the limiting case of an ultrafast process and incoherent time, equations (147), (149), (151) and (152) become

$$\xi_t^{ui} = \csc \beta_{uu}^{ui} \psi (\partial \theta_\psi / \partial t)^2 \quad (159)$$

$$\theta_{\xi t}^{ui} = \theta_\psi + \beta_{uu}^{ui} + \pi/2 \quad (160)$$

where from equations (47) and (145) it follows that

$$\tan \beta_{uu}^{ui} = E_{\psi t}^{ui} / F_{\psi t}^{ui} \quad (161)$$

where

$$E_{\psi t}^{ui} = (\partial \theta_\psi / \partial t)^2 \quad E_{\psi t}^{ui} \geq 0 \quad (162)$$

$$F_{\psi t}^{ui} = \partial^2 \theta_\psi / \partial t^2 \quad F_{\psi t}^{ui} \leq 0 \quad (163)$$

Equation (161) gives

$$\csc \beta_{uu}^{ui} = [(E_{\psi t}^{ui})^2 + (F_{\psi t}^{ui})^2]^{1/2} / E_{\psi t}^{ui} \quad (164)$$

which combined with equations (159) and (162) gives

$$\xi_t^{ui} = \psi [(E_{\psi t}^{ui})^2 + (F_{\psi t}^{ui})^2]^{1/2} \quad (165)$$

Equation (161) and the choice of signs in equations (162) and (163) give

$$\beta_{uu}^{ui} = \pi/2 + \kappa_{\psi t} \quad (166)$$

where the positive angle $\kappa_{\psi t}$ can also be defined by

$$\tan \kappa_{\psi t} = |F_{\psi t}^{ui}|/E_{\psi t}^{ui} \quad (167)$$

and equations (160) and (166) give

$$\theta_{\xi t}^{ui} = \theta_{\psi} + \kappa_{\psi t} + \pi \quad (168)$$

The alternate representation in equations (154) and (155) gives

$$\xi_t^{ui+} = -\psi [(E_{\psi t}^{ui})^2 + (F_{\psi t}^{ui})^2]^{1/2} \quad (169)$$

$$\theta_{\xi t}^{ui+} = \theta_{\psi} + \kappa_{\psi t}$$

so that $\theta_{\xi t}^{ui+}$ is a small angle.

Case 3. Slow Process and Coherent Time.

For this case the following conditions are valid

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \beta_{tt}^{\psi} = \pi/2 \quad (171)$$

and a general expression for the second derivative of the wave function with respect to time is required which will accept equation (171) as a limiting case. Equations (30), (138) and (144) give

$$\xi_t = \sec \beta_{uu} \sin \beta_{tt}^{\psi} t^{-1} \partial u / \partial \theta_t^{\psi} \quad (172)$$

$$= \sec \beta_{uu} \sin \beta_{tt}^{\psi} t^{-1} \partial / \partial \theta_t^{\psi} (\sec \beta_{\psi\psi} \sin \beta_{tt}^{\psi} t^{-1} \partial \psi / \partial \theta_t^{\psi}) \quad (173)$$

$$\theta_{\xi t} = \theta_{\psi} + \beta_{\psi\psi} + \beta_{uu} - 2(\theta_t^{\psi} + \beta_{tt}^{\psi}) \quad (174)$$

where β_{uu} is given by equation (30), (32) and (145). The alternative description introduced in this case is

$$\begin{aligned} \bar{\xi}_t &= \xi_t \exp(j\theta_{\xi t}) = \xi_t' \exp(j\theta_{\xi t}') \\ &= \partial^2 \bar{\psi} / \partial \bar{t}^2 = \partial \bar{u} / \partial \bar{t} \end{aligned} \quad (175)$$

where

$$\xi_t' = -\xi_t \quad (176)$$

$$\theta_{\xi t}' = \theta_{\xi t} + \pi \quad (177)$$

The limiting case of a slow process in coherent time is obtained from equation (171) which combined with equations (173) and (174) gives

$$\xi_t^{sc} = \sec \beta_{uu}^{sc} t^{-2} \partial^2 \Psi / \partial \theta_t^2 \quad (178)$$

$$\theta_{\xi t}^{sc} = \beta_{uu}^{sc} - 2\theta_t^\Psi - \pi. \quad (179)$$

Equations (50) and (145) give

$$\tan \beta_{uu}^{sc} = E_{\Psi t}^{sc} / F_{\Psi t}^{sc} \quad (180)$$

where

$$E_{\Psi t}^{sc} = -\partial \Psi / \partial \theta_t^\Psi \quad E_{\Psi t}^{sc} \geq 0 \quad (181)$$

$$F_{\Psi t}^{sc} = \partial^2 \Psi / \partial \theta_t^2 \quad F_{\Psi t}^{sc} \geq 0 \quad (182)$$

$$\sec \beta_{uu}^{sc} = [(E_{\Psi t}^{sc})^2 + (F_{\Psi t}^{sc})^2]^{1/2} / F_{\Psi t}^{sc} \quad (183)$$

so that β_{uu}^{sc} is a small positive angle. Equations (178), (182) and (183) give

$$\xi_t^{sc} = t^{-2} [(E_{\Psi t}^{sc})^2 + (F_{\Psi t}^{sc})^2]^{1/2} \quad (184)$$

Finally, the alternative description of the second derivative of the wave function with respect to time is obtained from equations (176), (177), (179) and (184) to be

$$\xi_t^{sc'} = -t^{-2} [(E_{\Psi t}^{sc})^2 + (F_{\Psi t}^{sc})^2]^{1/2} \quad (185)$$

$$\theta_{\xi t}^{sc'} = \beta_{uu}^{sc} - 2\theta_t^\Psi \quad (186)$$

where $\theta_{\xi t}^{sc'}$ is a small angle.

Case 4. Ultrafast Process and Coherent Time.

In this section a representation of the second derivative of the wave function with respect to time is determined which has the proper form to be used to obtain the limiting case of an ultrafast process with a coherent time variation that is described by

$$\beta_{\psi\psi} = \pi/2 \quad \beta_{tt}^{\psi} = \pi/2 \quad (187)$$

For this case the proper magnitude and internal phase angle of the second derivative is obtained from equations (31), (138) and (144) to be

$$\xi_t = \csc \beta_{uu} \sin \beta_{tt}^{\psi} u/t \partial \theta_u / \partial \theta_t^{\psi} \quad (188)$$

$$= \csc \beta_{uu} \sin^2 \beta_{tt}^{\psi} \csc \beta_{\psi\psi} \psi/t^2 \partial \theta_{\psi} / \partial \theta_t^{\psi} \partial \theta_u / \partial \theta_t^{\psi} \quad (189)$$

$$\theta_{\xi t} = \theta_{\psi} + \beta_{\psi\psi} + \beta_{uu} - 2(\theta_t^{\psi} + \beta_{tt}^{\psi}) \quad (190)$$

where β_{uu} is given by equations (31), (32) and (145), and where equation (65) gives

$$\partial \theta_u / \partial \theta_t^{\psi} = \partial \theta_{\psi} / \partial \theta_t^{\psi} - 1 + \partial / \partial \theta_t^{\psi} (\beta_{\psi\psi} - \beta_{tt}^{\psi}) \quad (191)$$

From equations (31) and (172) a second form of the second derivative of the wave function with respect to time can be written as

$$\xi_t = \sec \beta_{uu} \sin \beta_{tt}^{\psi} t^{-1} \partial u / \partial \theta_t^{\psi} \quad (192)$$

$$= \sec \beta_{uu} \sin \beta_{tt}^{\psi} t^{-1} \partial / \partial \theta_t^{\psi} (\csc \beta_{\psi\psi} \sin \beta_{tt}^{\psi} \psi/t \partial \theta_{\psi} / \partial \theta_t^{\psi}) \quad (193)$$

Combining equations (189) and (193) gives

$$\tan \beta_{uu} = C_{\psi t} / D_{\psi t} \quad (194)$$

$$\csc \beta_{uu} = (C_{\psi t}^2 + D_{\psi t}^2)^{1/2} / C_{\psi t} \quad (195)$$

$$\sec \beta_{uu} = (C_{\psi t}^2 + D_{\psi t}^2)^{1/2} / D_{\psi t} \quad (196)$$

where

$$C_{\psi t} = \sin \beta_{tt}^{\psi} \csc \beta_{\psi\psi} \psi/t \partial \theta_{\psi} / \partial \theta_t^{\psi} \partial \theta_u / \partial \theta_t^{\psi} \quad (197)$$

$$D_{\psi t} = d/d\theta_t^{\psi} (\csc \beta_{\psi\psi} \sin \beta_{tt}^{\psi} \psi/t \partial \theta_{\psi} / \partial \theta_t^{\psi}) \quad (198)$$

where $\partial \theta_u / \partial \theta_t^{\psi}$ is given by equation (191). The magnitude of the second derivative of the wave function with respect to time can then be written as

$$\xi_t = C_{\psi t} t^{-1} \csc \beta_{uu} \sin \beta_{tt}^{\psi} \quad (199)$$

$$= D_{\psi t} t^{-1} \sec \beta_{uu} \sin \beta_{tt}^{\psi} \quad (200)$$

$$= (C_{\psi t}^2 + D_{\psi t}^2)^{1/2} t^{-1} \sin \beta_{tt}^{\psi} \quad (201)$$

It is convenient for the case of an ultrafast process and coherent time to introduce the following alternative representation of the second derivative with respect to time

$$\begin{aligned}\bar{\xi}_t &= \xi_t \exp(j\theta_{\xi t}) = \xi'_t \exp(j\theta'_{\xi t}) \\ &= \partial^2 \bar{\Psi} / \partial \bar{t}^2 = \partial \bar{u} / \partial \bar{t}\end{aligned}\quad (202)$$

where

$$\xi'_t = -\xi_t \quad (203)$$

$$\theta'_{\xi t} = \theta_{\xi t} + \pi \quad (204)$$

In the limit of an ultrafast process and a coherent time variation, described by equation (187), it follows that equations (189), (193), (201) and (203) become

$$\xi_t^{uc} = -\csc \beta_{uu}^{uc} \Psi/t^2 E_{\Psi t}^{uc} \quad (205)$$

$$= -\sec \beta_{uu}^{uc} \Psi/t^2 F_{\Psi t}^{uc} \quad (206)$$

$$= -t^{-1} [(C_{\Psi t}^{uc})^2 + (D_{\Psi t}^{uc})^2]^{1/2} \quad (207)$$

$$= -\Psi t^{-2} [(E_{\Psi t}^{uc})^2 + (F_{\Psi t}^{uc})^2]^{1/2} \quad (208)$$

where

$$E_{\Psi t}^{uc} = \partial \theta_{\Psi} / \partial \theta_t^{\Psi} (\partial \theta_{\Psi} / \partial \theta_t^{\Psi} - 1) \quad E_{\Psi t}^{uc} \leq 0 \quad (209)$$

$$F_{\Psi t}^{uc} = \partial^2 \theta_{\Psi} / \partial \theta_t^{\Psi 2} \quad F_{\Psi t}^{uc} \geq 0 \quad (210)$$

$$C_{\Psi t}^{uc} = \Psi t^{-1} E_{\Psi t}^{uc} \quad (211)$$

$$D_{\Psi t}^{uc} = \Psi t^{-1} F_{\Psi t}^{uc} \quad (212)$$

From equations (47) and (145) or from equation (194) it follows that

$$\tan \beta_{uu}^{uc} = E_{\Psi t}^{uc} / F_{\Psi t}^{uc} \quad (213)$$

$$\csc \beta_{uu}^{uc} = (E_{\Psi t}^{uc})^{-1} [(E_{\Psi t}^{uc})^2 + (F_{\Psi t}^{uc})^2]^{1/2} \quad (214)$$

$$\sec \beta_{uu}^{uc} = (F_{\Psi t}^{uc})^{-1} [(E_{\Psi t}^{uc})^2 + (F_{\Psi t}^{uc})^2]^{1/2} \quad (215)$$

Combining equation (213) with the choice of signs in equations (209) and (210) gives

$$\beta_{uu}^{uc} = -\pi/2 + \delta_{\psi t} \quad (216)$$

where $\delta_{\psi t}$ is a positive angle given by

$$\tan \delta_{\psi t} = F_{\psi t}^{uc} / |E_{\psi t}^{uc}| \quad (217)$$

Equations (187), (190), (204) and (216) give

$$\theta_{\xi t}^{uc'} = \theta_{\psi} + \beta_{uu}^{uc} - 2\theta_t^{\psi} + \pi/2 \quad (218)$$

$$= \theta_{\psi} - 2\theta_t^{\psi} + \delta_{\psi t} \quad (219)$$

and therefore $\theta_{\xi t}^{uc'}$ is a small angle. Equations (208) and (219) are equivalent to the following complex number representation of the second derivative of the wave function with respect to time

$$\bar{\xi}_t^{uc} = (\partial^2 \bar{\psi} / \partial \bar{t}^2)^{uc} = \bar{\psi} / \bar{t}^2 [\partial \theta_{\psi} / \partial \theta_t^{\psi} (\partial \theta_{\psi} / \partial \theta_t^{\psi} - 1) - j \partial^2 \theta_{\psi} / \partial \theta_t^{\psi 2}] \quad (220)$$

which corresponds to an ultrafast process with a coherent time variation.

3. WAVE EQUATIONS IN BROKEN SYMMETRY SPACETIME. This section develops the various forms taken by the wave equation for slow and fast processes that can occur in coherent and incoherent space and coherent and incoherent time. In three dimensions the complex number wave equation is written as the following generalization of the standard scalar form of the wave equation in cartesian coordinates¹⁸⁻²⁴

$$\sum_{\alpha} \partial^2 \bar{\psi} / \partial \bar{\alpha}^2 = 1 / \bar{c}^2 \partial^2 \bar{\psi} / \partial \bar{t}^2 \quad (221)$$

where $\alpha = x, y, z$, and where the complex number space and time coordinates $\bar{\alpha}$ and \bar{t} are given by equations (1) and (2) respectively, and where

$$\bar{\psi} = \psi \exp(j\theta_{\psi}) \quad \bar{c} = c \exp(j\theta_c) \quad (222)$$

For wave propagation in a single space dimension equation (221) becomes

$$\partial^2 \bar{\psi} / \partial \bar{x}^2 = 1 / \bar{c}^2 \partial^2 \bar{\psi} / \partial \bar{t}^2 \quad (223)$$

From equations (55) and (138) it follows that the wave equation (221) can be written as

$$\sum_{\alpha} \bar{\xi}_{\alpha} = \bar{\xi}_t / \bar{c}^2 \quad (224)$$

or

$$\sum_{\alpha} \xi_{\alpha} \exp(j\theta_{\xi\alpha}) = c^{-2} \xi_t \exp[j(\theta_{\xi t} - 2\theta_c)] \quad (225)$$

For a one dimensional case equation (225) becomes

$$\xi_x \exp(j\theta_{\xi x}) = c^{-2} \xi_t \exp[j(\theta_{\xi t} - 2\theta_c)] \quad (226)$$

where $\theta_c = 0$ for the light speed in the vacuum.

An exact solution of the complex number wave equation (225) requires that the real and imaginary parts be taken, but this leads to very complicated equations. A simpler, but approximate, way of solving equation (225) is to make the assumption that the internal phase angles of each of the component terms of equation (225) are equal. However, because the phase angles on either side of equation (225) or (226) can have a zero or a $\pm \pi$ term included, as shown in Section 2 in equations (61), (72), (94), (121), (144), (155), (177) and (204), the solutions of equations (225) or (226) require that the eight possible process and spacetime states be considered individually, so it follows that

$$\pm \sum_{\alpha} \xi_{\alpha} = \pm c^{-2} \xi_t \quad (227)$$

$$\theta_{\xi\alpha} \pm (\pi, 0) = \theta_{\xi t} - 2\theta_c \pm (\pi, 0) \quad (228)$$

where $\alpha = x, y, z$ for the three dimensional case, and $\alpha = x$ for the one dimensional case. From the general equations (61) and (149) it is clear that equation (228) can be written as

$$\beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) \pm (\pi, 0) = \beta_{uu} - 2(\theta_t^{\psi} + \beta_{tt}^{\psi}) - 2\theta_c \pm (\pi, 0) \quad (229)$$

which gives the possible phase angle conditions.

The wave equation components given in equations (227) and (228) will now be investigated for eight possible states associated with the slow or ultrafast process rates, coherent or incoherent space coordinate variation, and the coherent or incoherent time coordinate variation.

Case a. Slow Process, Incoherent Space and Incoherent Time.

This section develops a general form of the wave equation that can be used to make the transition to the limiting case of a slow process occurring in incoherent space and incoherent time which is described by

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad \theta_t^{\psi} = 0 \quad \beta_{tt}^{\psi} = 0 \quad (230)$$

where $\alpha = x, y, z$. For the case considered all internal phase angles are small numbers so that the approximate solution to equation (225) can be obtained from equations (227) and (228) to be

$$\sum_{\alpha} \xi_{\alpha} = c^{-2} \xi_t \quad (231)$$

$$\theta_{\xi\alpha} = \theta_{\xi t} - 2\theta_c \quad (232)$$

where equation (232) is valid for $\alpha = x, y, z$. Equations (58), (141) and (231) give

$$\begin{aligned} & \sum_{\alpha} \sec \beta_{v\alpha v\alpha} \cos \beta_{\alpha\alpha} \partial/\partial\alpha (\sec \beta_{\psi\psi} \cos \beta_{\alpha\alpha} \partial\psi/\partial\alpha) \\ & = c^{-2} \sec \beta_{uu} \cos \beta_{tt}^{\psi} \partial/\partial t (\sec \beta_{\psi\psi} \cos \beta_{tt}^{\psi} \partial\psi/\partial t) \end{aligned} \quad (233)$$

For slowly changing values of $\beta_{\alpha\alpha}$, $\beta_{\psi\psi}$ and β_{tt}^{ψ} equation (233) becomes

$$\sum_{\alpha} \sec \beta_{v\alpha v\alpha} \cos^2 \beta_{\alpha\alpha} \partial^2 \psi / \partial \alpha^2 \sim c^{-2} \sec \beta_{uu} \cos^2 \beta_{tt}^{\psi} \partial^2 \psi / \partial t^2 \quad (234)$$

where $\beta_{v\alpha v\alpha}$ and β_{uu} are given by equations (62) and (145) respectively. From equation (229) it is clear that equation (232) can be written as

$$\beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) = \beta_{uu} - 2(\theta_t^{\psi} + \beta_{tt}^{\psi}) - 2\theta_c \quad (235)$$

for $\alpha = x, y, z$, where $\beta_{\alpha\alpha}$ and β_{tt}^{ψ} are defined in equations (13) and (14) respectively, and θ_{α} and θ_t are given in equations (1) and (2).

For the special case of a slow process in incoherent space and incoherent time described by equation (230), it follows that equations (233) or (234) reduce to the standard form of the wave equation

$$\sum_{\alpha} \partial^2 \psi / \partial \alpha^2 = c^{-2} \partial^2 \psi / \partial t^2 \quad (236)$$

For the one dimensional case equation (236) becomes

$$\partial^2 \psi / \partial x^2 = c^{-2} \partial^2 \psi / \partial t^2 \quad (237)$$

The phase angle equation (235) yields a trivial null result assuming $\theta_c = 0$ for a slow process in incoherent space and time.

Case b. Slow Process, Coherent Space and Incoherent Time.

In this section a general form of the wave equation is developed that can be used to pass to the limit of a slow process that occurs in coherent space and incoherent time whose description is given by

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \beta_{\alpha\alpha} = \pi/2 \quad \theta_t = 0 \quad \beta_{tt}^{\psi} = 0 \quad (238)$$

for $\alpha = x, y, z$. Equations (90), (92), (93), (141) and (227) become

$$-\sum_{\alpha} \xi_{\alpha} = c^{-2} \xi_t \quad (239)$$

or equivalently

$$\begin{aligned}
 & - \int_{\alpha} \sec \beta_{\alpha\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial/\partial\theta_{\alpha} (\sec \beta_{\psi\psi} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial\psi/\partial\theta_{\alpha}) \\
 & = c^{-2} \sec \beta_{uu} \cos \beta_{tt}^{\psi} \partial/\partial t (\sec \beta_{\psi\psi} \cos \beta_{tt}^{\psi} \partial\psi/\partial t)
 \end{aligned} \tag{240}$$

while equations (61), (92), (94) and (228) give

$$\theta'_{\xi\alpha} = \theta_{\xi t} - 2\theta_c \tag{241}$$

or

$$\theta_{\xi\alpha} + \pi = \theta_{\xi t} - 2\theta_c \tag{242}$$

which is equivalent to equation (229) written as

$$\beta_{\alpha\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) + \pi = \beta_{uu} - 2(\theta_t^{\psi} + \beta_{tt}^{\psi}) - 2\theta_c \tag{243}$$

For the special case of a slow process in coherent space and incoherent time as described by equation (238) it follows that equation (240) becomes

$$- \int_{\alpha} \sec \beta_{\alpha\alpha}^{\text{sc}} \alpha^{-2} \partial^2 \psi / \partial \theta_{\alpha}^2 = c^{-2} \partial^2 \psi / \partial t^2 \tag{244}$$

Combining equations (100) and (244) gives

$$- \int_{\alpha} \alpha^{-2} [(E_{\psi\alpha}^{\text{sc}})^2 + (F_{\psi\alpha}^{\text{sc}})^2]^{1/2} = c^{-2} \partial^2 \psi / \partial t^2 \tag{245}$$

where $E_{\psi\alpha}^{\text{sc}}$ and $F_{\psi\alpha}^{\text{sc}}$ are given by equations (98) and (99) respectively. For one space dimension equation (245) becomes

$$- c^2 x^{-2} [(E_{\psi x}^{\text{sc}})^2 + (F_{\psi x}^{\text{sc}})^2]^{1/2} = \partial^2 \psi / \partial t^2 \tag{246}$$

Combining equations (238) and (243) gives

$$\beta_{\alpha\alpha}^{\text{sc}} - 2\theta_{\alpha} = -2\theta_c \tag{247}$$

Combining equations (97) through (99) with equation (247) gives

$$\tan[2(\theta_{\alpha} - \theta_c)] = E_{\psi\alpha}^{\text{sc}} / F_{\psi\alpha}^{\text{sc}} \tag{248}$$

or

$$\tan[2(\theta_{\alpha} - \theta_c)] = - (\partial\psi/\partial\theta_{\alpha}) / (\partial^2 \psi / \partial \theta_{\alpha}^2) \tag{249}$$

which is valid for a slow process in coherent space and incoherent time.

Case c. Slow Process, Incoherent Space and Coherent Time.

This section derives a general form of the wave equation which can be used to attain the limiting case of a slow process occurring in incoherent space and coherent time which is defined by the following characteristics

$$\theta_{\psi} = 0 \quad \beta_{\psi\psi} = 0 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad \beta_{tt}^{\psi} = \pi/2 \quad (250)$$

Equations (58), (173), (176) and (227) give

$$\sum_{\alpha} \xi_{\alpha} = -c^{-2} \xi_t \quad (251)$$

or equivalently

$$\begin{aligned} & \sum_{\alpha} \sec \beta_{\alpha\alpha} \cos \beta_{\alpha\alpha} \partial/\partial\alpha (\sec \beta_{\psi\psi} \cos \beta_{\alpha\alpha} \partial\psi/\partial\alpha) \\ & = -c^{-2} \sec \beta_{uu} \sin \beta_{tt}^{\psi} t^{-1} \partial/\partial\theta_t^{\psi} (\sec \beta_{\psi\psi} \sin \beta_{tt}^{\psi} t^{-1} \partial\psi/\partial\theta_t^{\psi}) \end{aligned} \quad (252)$$

while equations (60), (174) and (177) give

$$\theta_{\xi\alpha} = \theta_{\xi t}^{\psi} - 2\theta_c \quad (253)$$

or

$$\theta_{\xi\alpha} = \theta_{\xi t}^{\psi} - 2\theta_c + \pi \quad (254)$$

or equivalently from equations (61), (174) and (229)

$$\beta_{\alpha\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) = \beta_{uu} - 2(\theta_t^{\psi} + \beta_{tt}^{\psi}) - 2\theta_c + \pi \quad (255)$$

For the limiting case of a slow wave propagation process in incoherent space and coherent time, equations (250) and (252) give

$$\sum_{\alpha} \partial^2 \psi / \partial \alpha^2 = -c^{-2} \sec \beta_{uu}^{\text{sc}} t^{-2} \partial^2 \psi / \partial \theta_t^{\psi 2} \quad (256)$$

Combining equations (182), (183) and (256) gives

$$\sum_{\alpha} \partial^2 \psi / \partial \alpha^2 = -c^{-2} t^{-2} [(E_{\psi t}^{\text{sc}})^2 + (F_{\psi t}^{\text{sc}})^2]^{1/2} \quad (257)$$

where $E_{\psi t}^{\text{sc}}$ and $F_{\psi t}^{\text{sc}}$ are given by equations (181) and (182). For one space dimension equation (257) becomes

$$c^2 t^2 \partial^2 \psi / \partial x^2 = -[(E_{\psi t}^{\text{sc}})^2 + (F_{\psi t}^{\text{sc}})^2]^{1/2} \quad (258)$$

Combining equations (250) and (255) gives the phase angle condition for this limiting case of the wave equation as

$$0 = \beta_{uu}^{sc} - 2\theta_t^\Psi - 2\theta_c \quad (259)$$

Combining equations (180) through (182) and (259) gives the phase angle condition as

$$\tan[2(\theta_t^\Psi + \theta_c)] = E_{\Psi t}^{sc}/F_{\Psi t}^{sc} \quad (260)$$

or equivalently

$$\tan[2(\theta_t^\Psi + \theta_c)] = - (\partial\Psi/\partial\theta_t^\Psi)/(\partial^2\Psi/\partial\theta_t^{\Psi 2}) \quad (261)$$

Case d. Slow Process, Coherent Space and Coherent Time.

A general form of the wave equation is derived in this section that can be used to obtain the limiting case of a slow process occurring in coherent space and coherent time which is described by

$$\theta_\Psi = 0 \quad \beta_{\Psi\Psi} = 0 \quad \beta_{\alpha\alpha} = \pi/2 \quad \beta_{tt}^\Psi = \pi/2 \quad (262)$$

The combination of equations (90), (93), (173) and (227) gives

$$-\sum_{\alpha} \xi_{\alpha} = -c^{-2} \xi_t \quad (263)$$

or equivalently

$$\begin{aligned} & \sum_{\alpha} \sec \beta_{v\alpha v\alpha} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial/\partial\theta_{\alpha} (\sec \beta_{\Psi\Psi} \sin \beta_{\alpha\alpha} \alpha^{-1} \partial\Psi/\partial\theta_{\alpha}) \\ & = c^{-2} \sec \beta_{uu} \sin \beta_{tt}^\Psi t^{-1} \partial/\partial\theta_t^\Psi (\sec \beta_{\Psi\Psi} \sin \beta_{tt}^\Psi t^{-1} \partial\Psi/\partial\theta_t^\Psi) \end{aligned} \quad (264)$$

while equations (60), (94), (174), (177) and (229) give

$$\theta'_{\xi\alpha} = \theta'_{\xi t} - 2\theta_c \quad (265)$$

or

$$\beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) = \beta_{uu} - 2(\theta_t^\Psi + \beta_{tt}^\Psi) - 2\theta_c \quad (266)$$

where $+\pi$ has been cancelled from both sides of equation (266).

For the special case of a slow wave propagation process in coherent space and coherent time, equations (262) and (264) give

$$\sum_{\alpha} \sec \beta_{v\alpha v\alpha}^{sc} \alpha^{-2} \partial^2\Psi/\partial\theta_{\alpha}^2 = c^{-2} \sec \beta_{uu}^{sc} t^{-2} \partial^2\Psi/\partial\theta_t^{\Psi 2} \quad (267)$$

Combining equations (99), (100), (182), (183) and (267) gives

$$\sum_{\alpha} \alpha^{-2} [(E_{\Psi\alpha}^{sc})^2 + (F_{\Psi\alpha}^{sc})^2]^{1/2} = c^{-2} t^{-2} [(E_{\Psi t}^{sc})^2 + (F_{\Psi t}^{sc})^2]^{1/2} \quad (268)$$

where $E_{\psi\alpha}^{sc}$, $F_{\psi\alpha}^{sc}$, $E_{\psi t}^{sc}$ and $F_{\psi t}^{sc}$ are given by equations (98), (99), (181) and (182) respectively. For one space dimension, equation (268) becomes

$$(ct/x)^2 [(E_{\psi x}^{sc})^2 + (F_{\psi x}^{sc})^2]^{1/2} = [(E_{\psi t}^{sc})^2 + (F_{\psi t}^{sc})^2]^{1/2} \quad (269)$$

Combining equations (262) and (266) gives

$$\beta_{v\alpha v\alpha}^{sc} - 2\theta_{\alpha} = \beta_{uu}^{sc} - 2\theta_t^{\psi} - 2\theta_c \quad (270)$$

Equation (270) can be rewritten using

$$\tan^{-1}(E_{\psi\alpha}^{sc}/F_{\psi\alpha}^{sc}) - \tan^{-1}(E_{\psi t}^{sc}/F_{\psi t}^{sc}) = 2(\theta_{\alpha} - \theta_t^{\psi} - \theta_c) \quad (271)$$

which is valid for a slow process in coherent space and coherent time.

Case e. Ultrafast Process, Incoherent Space and Incoherent Time.

This section develops the general form of the wave equation that can be utilized to attain the limiting case of an ultrafast process in incoherent space and incoherent time which is represented by

$$\beta_{\psi\psi} = \pi/2 \quad \theta_{\alpha} = 0 \quad \beta_{\alpha\alpha} = 0 \quad \theta_t^{\psi} = 0 \quad \beta_{tt}^{\psi} = 0 \quad (272)$$

Combining equations (66), (71), (149), (154) and (227) gives

$$-\sum_{\alpha} \xi_{\alpha} = -c^{-2} \xi_t \quad (273)$$

or equivalently

$$\begin{aligned} \sum_{\alpha} \csc \beta_{v\alpha v\alpha} \cos^2 \beta_{\alpha\alpha} \partial \theta_{\psi} / \partial \alpha \partial \theta_{v\alpha} / \partial \alpha \\ = c^{-2} \csc \beta_{uu} \cos^2 \beta_{tt}^{\psi} \partial \theta_{\psi} / \partial t \partial \theta_u / \partial t \end{aligned} \quad (274)$$

where a factor $\psi \csc \beta_{\psi\psi}$ has been divided out of both sides of equation (274). The phase angle condition is obtained from equations (69), (72), (151), (155) and (229) as

$$\theta_{\xi\alpha}^{\dagger} = \theta_{\xi t}^{\dagger} - 2\theta_c \quad (275)$$

or equivalently

$$\beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) = \beta_{uu} - 2(\theta_t^{\psi} + \beta_{tt}^{\psi}) - 2\theta_c \quad (276)$$

where $-\pi$ has been cancelled from both sides of equation (276).

For the special case of an ultrafast wave propagation process in incoherent space and incoherent time, equations (36), (47), (272) and (274) give

$$\sum_{\alpha} \csc \beta_{\nu\alpha\nu\alpha}^{ui} (\partial\theta_{\Psi}/\partial\alpha)^2 = c^{-2} \csc \beta_{uu}^{ui} (\partial\theta_{\Psi}/\partial t)^2 \quad (277)$$

From equations (79), (81) (162), (164) and (277) it follows that

$$\sum_{\alpha} [(E_{\Psi\alpha}^{ui})^2 + (F_{\Psi\alpha}^{ui})^2]^{1/2} = c^{-2} [(E_{\Psi t}^{ui})^2 + (F_{\Psi t}^{ui})^2]^{1/2} \quad (278)$$

where $E_{\Psi\alpha}^{ui}$, $F_{\Psi\alpha}^{ui}$, $E_{\Psi t}^{ui}$ and $F_{\Psi t}^{ui}$ are given by equations (79), (80), (162) and (163) respectively. For one space dimension equation (278) becomes

$$c^2 [(E_{\Psi x}^{ui})^2 + (F_{\Psi x}^{ui})^2]^{1/2} = [(E_{\Psi t}^{ui})^2 + (F_{\Psi t}^{ui})^2]^{1/2} \quad (279)$$

Combining equations (272) and (276) gives the phase angle equation for the wave equation in this special case as

$$\beta_{\nu\alpha\nu\alpha}^{ui} = \beta_{uu}^{ui} - 2\theta_c \quad (280)$$

and finally using equations (83), (166) and (280) gives

$$\kappa_{\Psi\alpha} = \kappa_{\Psi t} - 2\theta_c \quad (281)$$

where $\kappa_{\Psi\alpha}$ and $\kappa_{\Psi t}$ are given by equations (84) and (167) respectively.

Case f. Ultrafast Process, Coherent Space and Incoherent Time.

Consider now the development of a general form of the wave equation that can be used to obtain the limiting case of an ultrafast wave propagation process in coherent space and incoherent time which is described by

$$\beta_{\Psi\Psi} = \pi/2 \quad \beta_{\alpha\alpha} = \pi/2 \quad \theta_t^{\Psi} = 0 \quad \beta_{tt}^{\Psi} = 0 \quad (282)$$

Combining equations (106), (120), (149), (154) and (227) gives

$$-\sum_{\alpha} \xi_{\alpha} = -c^{-2} \xi_t \quad (283)$$

or equivalently

$$\begin{aligned} \sum_{\alpha} \csc \beta_{\nu\alpha\nu\alpha} \sin^2 \beta_{\alpha\alpha} \alpha^{-2} \partial\theta_{\Psi}/\partial\theta_{\alpha} \partial\theta_{\nu\alpha}/\partial\theta_{\alpha} \\ = c^{-2} \csc \beta_{uu} \cos^2 \beta_{tt}^{\Psi} \partial\theta_{\Psi}/\partial t \partial\theta_u/\partial t \end{aligned} \quad (284)$$

where a factor $\Psi \csc \beta_{\Psi\Psi}$ has been divided out of both sides of equation (284). Combining equations (107), (121), (151), (155), (228) and (229) gives

$$\theta_{\xi\alpha}^{\dagger} = \theta_{\xi t}^{\dagger} - 2\theta_c \quad (285)$$

or

$$\theta_{\xi\alpha} + \pi = \theta_{\xi t} - 2\theta_c - \pi \quad (286)$$

or equivalently

$$\beta_{\nu\alpha\nu\alpha} - 2(\theta_\alpha + \beta_{\alpha\alpha}) + \pi = \beta_{uu} - 2(\theta_t^\Psi + \beta_{tt}^\Psi) - 2\theta_c - \pi \quad (287)$$

For the special case of an ultrafast wave propagation process in coherent space and incoherent time, equations (42), (47), (282) and (284) give

$$\sum_\alpha \csc \beta_{\nu\alpha\nu\alpha}^{\text{uc}} \alpha^{-2} \partial\theta_\Psi / \partial\theta_\alpha (\partial\theta_\Psi / \partial\theta_\alpha - 1) = c^{-2} \csc \beta_{uu}^{\text{ui}} (\partial\theta_\Psi / \partial t)^2 \quad (288)$$

Combining equations (126), (131), (162), (164) and (288) gives

$$\sum_\alpha \alpha^{-2} [(E_{\Psi\alpha}^{\text{uc}})^2 + (F_{\Psi\alpha}^{\text{uc}})^2]^{1/2} = c^{-2} [(E_{\Psi t}^{\text{ui}})^2 + (F_{\Psi t}^{\text{ui}})^2]^{1/2} \quad (289)$$

where $E_{\Psi\alpha}^{\text{uc}}$, $F_{\Psi\alpha}^{\text{uc}}$, $E_{\Psi t}^{\text{ui}}$ and $F_{\Psi t}^{\text{ui}}$ are given by equations (126), (127), (162) and (163) respectively. For one space dimension equation (289) becomes

$$c^2 x^{-2} [(E_{\Psi x}^{\text{uc}})^2 + (F_{\Psi x}^{\text{uc}})^2]^{1/2} = [(E_{\Psi t}^{\text{ui}})^2 + (F_{\Psi t}^{\text{ui}})^2]^{1/2} \quad (290)$$

From equations (282) and (287) it follows that

$$\beta_{\nu\alpha\nu\alpha}^{\text{uc}} - 2\theta_\alpha = \beta_{uu}^{\text{ui}} - 2\theta_c - \pi \quad (291)$$

Combining equations (133), (166) and (291) gives

$$\delta_{\Psi\alpha} - 2\theta_\alpha = \kappa_{\Psi t} - 2\theta_c \quad (292)$$

where $\delta_{\Psi\alpha}$ and $\kappa_{\Psi t}$ are defined by equations (134) and (167) respectively.

Case g. Ultrafast Process, Incoherent Space and Coherent Time.

This section gives the general wave equation for spacetime with broken internal symmetries that can be used to approach the limiting case of an ultrafast wave propagation process in incoherent space and coherent time whose characteristics are given by

$$\beta_{\Psi\Psi} = \pi/2 \quad \theta_\alpha = 0 \quad \beta_{\alpha\alpha} = 0 \quad \beta_{tt}^\Psi = \pi/2 \quad (293)$$

From equations (66), (71), (189), (203) and (227) it follows that

$$-\sum_\alpha \xi_\alpha = -c^{-2} \xi_t \quad (294)$$

or

$$\begin{aligned} & \sum_\alpha \csc \beta_{\nu\alpha\nu\alpha} \cos^2 \beta_{\alpha\alpha} \partial\theta_\Psi / \partial\alpha \partial\theta_{\nu\alpha} / \partial\alpha \\ & = c^{-2} \csc \beta_{uu} \sin^2 \beta_{tt}^\Psi t^{-2} \partial\theta_\Psi / \partial\theta_t^\Psi \partial\theta_u / \partial\theta_t^\Psi. \end{aligned} \quad (295)$$

where a factor $\Psi \csc \beta_{\Psi\Psi}$ has been divided out of both sides of equation (295). Combining equations (69), (72), (190), (204), (228) and (229) gives

$$\theta_{\xi\alpha}^+ = \theta_{\xi t}^+ - 2\theta_c \quad (296)$$

or

$$\theta_{\xi\alpha} - \pi = \theta_{\xi t} - 2\theta_c + \pi \quad (297)$$

or equivalently

$$\beta_{\Psi\Psi} - 2(\theta_{\alpha}^+ + \beta_{\alpha\alpha}) - \pi = \beta_{uu} - 2(\theta_t^{\Psi} + \beta_{tt}^{\Psi}) - 2\theta_c + \pi \quad (298)$$

For the special case of an ultrafast wave propagation process in incoherent space and coherent time, equations (36), (53), (293) and (295) give

$$\sum_{\alpha} \csc \beta_{\Psi\Psi}^{ui} (\partial\theta_{\Psi}/\partial\alpha)^2 = c^{-2} t^{-2} \csc \beta_{uu}^{uc} \partial\theta_{\Psi}/\partial\theta_t^{\Psi} (\partial\theta_{\Psi}/\partial\theta_t^{\Psi} - 1) \quad (299)$$

Combining equations (79), (81), (209), (214) and (299) gives

$$\sum_{\alpha} [(E_{\Psi\alpha}^{ui})^2 + (F_{\Psi\alpha}^{ui})^2]^{1/2} = c^{-2} t^{-2} [(E_{\Psi t}^{uc})^2 + (F_{\Psi t}^{uc})^2]^{1/2} \quad (300)$$

where $E_{\Psi\alpha}^{ui}$, $F_{\Psi\alpha}^{ui}$, $E_{\Psi t}^{uc}$ and $F_{\Psi t}^{uc}$ are given by equations (79), (80), (209) and (210) respectively. For one space dimension equation (300) becomes

$$c^2 t^2 [(E_{\Psi x}^{ui})^2 + (F_{\Psi x}^{ui})^2]^{1/2} = [(E_{\Psi t}^{uc})^2 + (F_{\Psi t}^{uc})^2]^{1/2} \quad (301)$$

Equations (293) and (298) give the internal phase angle equation for this special case of the wave equation as

$$\beta_{\Psi\Psi}^{ui} - \pi = \beta_{uu}^{uc} - 2\theta_t^{\Psi} - 2\theta_c \quad (302)$$

Combining equations (83), (216) and (302) gives

$$\kappa_{\Psi\alpha} = \delta_{\Psi t} - 2\theta_t^{\Psi} - 2\theta_c \quad (303)$$

where $\kappa_{\Psi\alpha}$ and $\delta_{\Psi t}$ are defined by equations (84) and (217) respectively.

Case h. Ultrafast Process, Coherent Space and Coherent Time.

In this section a completely general form of the wave equation in broken symmetry spacetime is developed for the purpose of attaining the limiting case of an ultrafast wave propagation process in coherent space and coherent time which is described by

$$\beta_{\Psi\Psi} = \pi/2 \quad \beta_{\alpha\alpha} = \pi/2 \quad \beta_{tt}^{\Psi} = \pi/2 \quad (304)$$

From equations (106), (120), (189), (203) and (227) it follows that the scalar

component of the wave equation is

$$-\sum_{\alpha} \xi_{\alpha} = -c^{-2} \xi_t \quad (305)$$

or equivalently

$$\begin{aligned} \sum_{\alpha} \csc \beta_{v\alpha v\alpha} \sin^2 \beta_{\alpha\alpha} \alpha^{-2} \partial \theta_{\Psi} / \partial \theta_{\alpha} \partial \theta_{v\alpha} / \partial \theta_{\alpha} \\ = c^{-2} \csc \beta_{uu} \sin^2 \beta_{tt}^{\Psi} t^{-2} \partial \theta_{\Psi} / \partial \theta_t^{\Psi} \partial \theta_u / \partial \theta_t^{\Psi} \end{aligned} \quad (306)$$

where a factor $\Psi \csc \beta_{\Psi\Psi}$ has been divided out of both sides of equation (306). Combining equations (107), (121), (190), (204), (228) and (229) gives the phase angle condition for the wave equation as

$$\theta'_{\xi\alpha} = \theta'_{\xi t} - 2\theta_c \quad (307)$$

or

$$\theta_{\xi\alpha} + \pi = \theta_{\xi t} - 2\theta_c + \pi \quad (308)$$

or equivalently as

$$\beta_{v\alpha v\alpha} - 2(\theta_{\alpha} + \beta_{\alpha\alpha}) = \beta_{uu} - 2(\theta_t^{\Psi} + \beta_{tt}^{\Psi}) - 2\theta_c \quad (309)$$

where $+\pi$ has been cancelled from both sides of equation (309).

In the special case of an ultrafast wave propagation process in coherent space and coherent time, equations (42), (53), (304) and (306) give

$$\begin{aligned} \sum_{\alpha} \csc \beta_{v\alpha v\alpha}^{\text{uc}} \alpha^{-2} \partial \theta_{\Psi} / \partial \theta_{\alpha} (\partial \theta_{\Psi} / \partial \theta_{\alpha} - 1) \\ = c^{-2} \csc \beta_{uu}^{\text{uc}} t^{-2} \partial \theta_{\Psi} / \partial \theta_t^{\Psi} (\partial \theta_{\Psi} / \partial \theta_t^{\Psi} - 1) \end{aligned} \quad (310)$$

Combining equations (126), (131), (209), (214) and (310) gives

$$\sum_{\alpha} \alpha^{-2} [(E_{\Psi\alpha}^{\text{uc}})^2 + (F_{\Psi\alpha}^{\text{uc}})^2]^{1/2} = (ct)^{-2} [(E_{\Psi t}^{\text{uc}})^2 + (F_{\Psi t}^{\text{uc}})^2]^{1/2} \quad (311)$$

where $E_{\Psi\alpha}^{\text{uc}}$, $F_{\Psi\alpha}^{\text{uc}}$, $E_{\Psi t}^{\text{uc}}$ and $F_{\Psi t}^{\text{uc}}$ are given by equations (126), (127), (209) and (210) respectively. For one space dimension equation (311) becomes

$$(ct/x)^2 [(E_{\Psi x}^{\text{uc}})^2 + (F_{\Psi x}^{\text{uc}})^2]^{1/2} = [(E_{\Psi t}^{\text{uc}})^2 + (F_{\Psi t}^{\text{uc}})^2]^{1/2} \quad (312)$$

From equations (304) and (309) it follows that the phase angle relationship for this special case of the wave equation is given by

$$\beta_{v\alpha v\alpha}^{\text{uc}} - 2\theta_{\alpha} = \beta_{uu}^{\text{uc}} - 2\theta_t^{\Psi} - 2\theta_c \quad (313)$$

Combining equations (133), (216) and (313) gives

$$\delta_{\Psi\alpha} - 2\theta_{\alpha} = \delta_{\Psi t} - 2\theta_t^{\Psi} - 2\theta_c \quad (314)$$

where $\delta_{\Psi\alpha}$ and $\delta_{\Psi t}$ are defined by equations (134) and (217) respectively. The results given above can also be obtained from equations (137) and (220) which give the wave equation for an ultrafast process in coherent space and coherent time as

$$\sum_{\alpha} \bar{\xi}_{\alpha}^{uc} = 1/\bar{c}^2 \bar{\xi}_t^{uc} \quad (315)$$

or equivalently

$$\begin{aligned} & \sum_{\alpha} 1/\bar{\alpha}^2 [\partial\theta_{\Psi}/\partial\theta_{\alpha} (\partial\theta_{\Psi}/\partial\theta_{\alpha} - 1) - j\partial^2\theta_{\Psi}/\partial\theta_{\alpha}^2] \\ & = (\bar{c}\bar{t})^{-2} [\partial\theta_{\Psi}/\partial\theta_t^{\Psi} (\partial\theta_{\Psi}/\partial\theta_t^{\Psi} - 1) - j\partial^2\theta_{\Psi}/\partial\theta_t^{\Psi 2}] \end{aligned} \quad (316)$$

Equation (316) is equivalent to equations (311) and (313). The one dimensional analog of equation (316) is given by

$$\begin{aligned} & (\bar{c}\bar{t}/\bar{x})^2 [\partial\theta_{\Psi}/\partial\theta_x (\partial\theta_{\Psi}/\partial\theta_x - 1) - j\partial^2\theta_{\Psi}/\partial\theta_x^2] \\ & = \partial\theta_{\Psi}/\partial\theta_t^{\Psi} (\partial\theta_{\Psi}/\partial\theta_t^{\Psi} - 1) - j\partial^2\theta_{\Psi}/\partial\theta_t^{\Psi 2} \end{aligned} \quad (317)$$

4. WAVE EQUATION SOLUTIONS. The general forms of the wave equations derived in Section 3 are too complicated to allow simple analytical solutions. This is true for equations (233), (240), (252), (264), (274), (284), (295) and (306) and their corresponding phase angle equations (232), (243), (255), (266), (276), (287), (298) and (309). For some special cases of the wave equation it is relatively easy to find analytical solutions, and it is these cases that are considered in this section.

Case a. Slow Process, Incoherent Space and Incoherent Time.

The standard wave equation (236) has the following well known solution¹⁸⁻²⁴

$$\Psi = A \exp[i(\sum_{\alpha} k_{\alpha} \alpha \pm \omega t)] \quad (318)$$

where $\alpha = x, y$ and z . For this case α and t are variables, and equation (318) is a solution to equation (236) only if the external space wave numbers k_{α} are related to the external frequency ω by the following standard equation

$$\sum_{\alpha} k_{\alpha}^2 = \omega^2/c^2 \quad (319)$$

where c = propagation speed. For the one dimensional case the solution to the wave equation is

$$\Psi = A \exp[ik_x(x \pm ct)] \quad (320)$$

where

$$k_x = \omega/c \quad (321)$$

The \pm signs in equations (318) and (320) correspond to the two possible directions for the wave propagation along each axis.

Case b. Slow Process, Coherent Space and Incoherent Time.

For this special case the wave equation (245) has the following simple solution

$$\Psi = A \exp\left(-\sum_{\alpha} a_{\alpha} \theta_{\alpha}\right) \exp\left[i\left(\sum_{\alpha} k_{\alpha} \alpha \pm \omega t\right)\right] \quad (322)$$

where θ_{α} and t are variables, $\alpha = x, y$ and $z = \text{constants}$, $a_{\alpha} = \text{constants}$, and where as usual k_{α} and ω are constants. From equations (98), (99) and (322) it follows that

$$E_{\Psi\alpha}^{\text{sc}} = a_{\alpha} \Psi \quad F_{\Psi\alpha}^{\text{sc}} = a_{\alpha}^2 \Psi \quad (323)$$

$$\partial^2 \Psi / \partial t^2 = -\omega^2 \Psi \quad (324)$$

Equations (98) and (323) give the condition $a_{\alpha} \geq 0$. Combining equations (245), (323) and (324) gives the condition between the internal space parameters a_{α} and the external time angular frequency ω required for a solution

$$\sum_{\alpha} \alpha^{-2} a_{\alpha} (a_{\alpha}^2 + 1)^{1/2} = (\omega/c)^2 \quad (325)$$

From equation (97) it follows that for this solution

$$\tan \beta_{\nu\alpha\nu}^{\text{sc}} = a_{\alpha}^{-1} \quad (326)$$

For the one dimensional case the wave function is

$$\Psi = A \exp(-a_x \theta_x) \exp[ik_x(x \pm ct)] \quad (327)$$

which gives the condition for a solution as

$$a_x (a_x^2 + 1)^{1/2} = (\omega x/c)^2 \quad (328)$$

which gives the constant a_x as

$$a_x = \left\{ -1/2 + 1/2[1 + 4(\omega x/c)^4]^{1/2} \right\}^{1/2} \quad (329)$$

where $a_x > 0$. This case corresponds to the wave amplitude Ψ propagating evanescently in internal space θ_{α} for fixed spatial coordinate magnitudes α , and oscillating in external time. For this special case the measured space coordinates change in time according to

$$\alpha_m = \alpha \cos \theta_\alpha \quad d\alpha_m/dt = -\alpha \sin \theta_\alpha d\theta_\alpha/dt \quad (330)$$

There is no wave propagation in external space for this case because $\alpha = \text{constant}$ in the phase factor of equation (322).

Case c. Slow Process, Incoherent Space and Coherent Time.

The wave equation corresponding to this special case is given by equation (257) and has the solution

$$\Psi = A \exp(-b_t \theta_t^\Psi) \exp[i(\sum_\alpha k_\alpha \alpha \pm \omega t)] \quad (331)$$

where α and θ_t^Ψ are variables, $t = \text{constant}$ and $b_t = \text{constant}$. Equations (181), (182) and (331) give

$$E_{\Psi t}^{sc} = b_t \Psi \quad F_{\Psi t}^{sc} = b_t^2 \Psi \quad (332)$$

$$\partial^2 \Psi / \partial \alpha^2 = -k_\alpha^2 \Psi$$

Equations (181) and (332) give $b_t > 0$. Combining equations (257), (332) and (333) gives the following connection between the external space wave numbers k_α and the internal time parameter b_t

$$\sum_\alpha k_\alpha^2 = (ct)^{-2} b_t^2 (b_t^2 + 1)^{1/2} \quad (334)$$

where $\alpha = x, y, z$. From equation (334) it follows immediately that for $b_t > 0$

$$b_t = \{ -1/2 + 1/2[1 + 4(ct)^4 (\sum_\alpha k_\alpha^2)^2]^{1/2} \}^{1/2} \quad (334A)$$

For this solution equation (180) gives the following phase angle condition

$$\tan \beta_{uu}^{sc} = b_t^{-1} \quad (335)$$

For the one space dimension case the solution to the wave equation is obtained from equation (331) as

$$\Psi = A \exp(-b_t \theta_t^\Psi) \exp[ik_x(x \pm ct)] \quad (336)$$

subject to

$$b_t(b_t^2 + 1)^{1/2} = (ctk_x)^2 \quad (337)$$

from which the constant b_t is determined to have the value

$$b_t = \{ -1/2 + 1/2[1 + 4(ctk_x)^4]^{1/2} \}^{1/2} \quad (338)$$

with $b_t > 0$. The wave function in equation (331) corresponds to $t = \text{constant}$

so that the wave propagates in external space with an evanescent wave propagation in internal time. However, the measured time does change during the motion because

$$t_m = t \cos \theta_t^\Psi \quad dt_m/dx = -t \sin \theta_t^\Psi d\theta_t^\Psi/dx \quad (339)$$

There is no wave propagation in external time because $t = \text{constant}$ in the phase angle of equation (331).

Case d. Slow Process, Coherent Space and Coherent Time.

For the special case the wave equation (268) has the solution

$$\Psi = A \exp(-\sum_{\alpha} d_{\alpha} \theta_{\alpha} - d_t \theta_t^\Psi) \exp[i(\sum_{\alpha} k_{\alpha} \alpha \pm \omega t)] \quad (340)$$

where θ_{α} and θ_t^Ψ are variables, α and $t = \text{constants}$, and where d_{α} and $d_t = \text{constants}$. Equations (98), (99), (181), (182), (268) and (340) give

$$E_{\Psi\alpha}^{SC} = d_{\alpha} \Psi \quad F_{\Psi\alpha}^{SC} = d_{\alpha}^2 \Psi \quad (341)$$

$$E_{\Psi t}^{SC} = d_t \Psi \quad F_{\Psi t}^{SC} = d_t^2 \Psi \quad (342)$$

Combining equations (98), (181), (341) and (342) gives $d_{\alpha} \geq 0$ and $d_t \geq 0$. Then combining equations (268), (341) and (342) gives the condition between d_{α} and d_t that is required for equation (340) to be a solution of the wave equation (268) as follows

$$\sum_{\alpha} \alpha^{-2} d_{\alpha} (d_{\alpha}^2 + 1)^{1/2} = (ct)^{-2} d_t (d_t^2 + 1)^{1/2} \quad (343)$$

or equivalently

$$d_t^2 = -1/2 + 1/2 \{1 + 4(ct)^4 [\sum_{\alpha} \alpha^{-2} d_{\alpha} (d_{\alpha}^2 + 1)^{1/2}]^2\}^{1/2} \quad (344)$$

where $d_t \geq 0$. For this solution the phase angle conditions in equations (97) and (180) give

$$\tan \beta_{v\alpha v\alpha}^{SC} = d_{\alpha}^{-1} \quad \tan \beta_{uu}^{SC} = d_t^{-1} \quad (345)$$

For the case of one space dimension the solution to equation (269) is written as

$$\Psi = A \exp(-d_x \theta_x - d_t \theta_t^\Psi) \exp[i(k_x x \pm \omega t)] \quad (346)$$

subject to the following condition between the constants d_x and d_t

$$(ct/x)^2 d_x (d_x^2 + 1)^{1/2} = d_t (d_t^2 + 1)^{1/2} \quad (347)$$

or equivalently

$$d_t = \{ -1/2 + 1/2[1 + 4(ct/x)^4 d_x^2 (d_x^2 + 1)]^{1/2} \}^{1/2} \quad (348)$$

where $d_x \geq 0$ and $d_t \geq 0$. Because $\alpha = \text{constant}$ and $t = \text{constant}$ there is no wave propagation in external space or time. This case corresponds to an evanescent propagation of the wave amplitude ψ in both internal space and internal time. However, the measured space coordinates α_m and the measured time coordinate t_m are changing during this process as shown by equations (329) and (338)

$$d\alpha_m/d\theta_t^\psi = -\alpha \sin \theta_\alpha d\theta_\alpha/d\theta_t^\psi \quad (349)$$

$$dt_m/d\theta_t^\psi = -t \sin \theta_t^\psi \quad (350)$$

Therefore this case should be physically observable.

Case e. Ultrafast Process, Incoherent Space and Incoherent Time.

For this special case a simple solution to equation (278) is written for the case of three dimensional cartesian coordinates as

$$\bar{\psi} = A \exp(j\theta_\psi) \quad (351)$$

$$\theta_\psi = \sum_{\alpha} e_{\alpha} \alpha + e_t t + e_o \quad (352)$$

where α and t are variables with $\alpha = x, y, z$, and where e_{α} , e_t and $e_o = \text{constants}$. In this case, equations (79), (80), (162) and (163) give

$$E_{\psi\alpha}^{ui} = e_{\alpha}^2 \quad F_{\psi\alpha}^{ui} = 0 \quad (353)$$

$$E_{\psi t}^{ui} = e_t^2 \quad F_{\psi t}^{ui} = 0 \quad (354)$$

where e_{α} and e_t are positive or negative real numbers. Combining equations (278), (353) and (354) gives

$$\sum_{\alpha} e_{\alpha}^2 = c^{-2} e_t^2 \quad e_t = \pm c \left(\sum_{\alpha} e_{\alpha}^2 \right)^{1/2} \quad (355)$$

which gives the relation between the constants e_{α} and e_t that is required in order for equations (351) and (352) to be a solution to the wave equation (278). For this solution equations (78) and (161) give the phase angle conditions as

$$\beta_{v\alpha\alpha}^{ui} = \pi/2 \quad \beta_{uu}^{ui} = \pi/2 \quad (356)$$

For the case of one space dimension the wave function is

$$\psi = A \exp(j\theta_\psi) \quad (357)$$

$$\theta_\psi = e_x x + e_t t + e_o \quad (358)$$

subject to the condition

$$e_t^2 = c^2 e_x^2 \quad e_t = \pm c e_x \quad (359)$$

so that

$$\theta_\psi = e_x (x \pm ct) \quad (360)$$

where the constant e_0 can always be absorbed into the multiplicative constant, so that finally

$$\bar{\Psi} = \bar{A} \exp[je_x (x \pm ct)] \quad (361)$$

where x and t are variables. This special case corresponds to the propagation of the internal phase angle θ_ψ in external space and time.

Case f. Ultrafast Process, Coherent Space and Incoherent Time.

In this special case a simple solution to equation (289) is

$$\Psi = A \exp(j\theta_\psi) \quad (362)$$

$$\theta_\psi = \sum_{\alpha} f_{\alpha} \theta_{\alpha} + f_t t + f_o \quad (363)$$

where θ_{α} and t are variables, $\alpha = x, y, z = \text{constants}$, and f_{α} , f_t and $f_o = \text{constants}$. For this case equations (126), (127), (162) and (163) give

$$E_{\psi\alpha}^{uc} = f_{\alpha} (f_{\alpha} - 1) \quad F_{\psi\alpha}^{uc} = 0 \quad (364)$$

$$E_{\psi t}^{ui} = f_t^2 \quad F_{\psi t}^{ui} = 0 \quad (365)$$

From equations (126) and (364) it follows that $E_{\psi\alpha}^{uc} \leq 0$ and $0 \leq f_{\alpha} \leq 1$. Combining equations (289), (364) and (365) gives

$$\sum_{\alpha} \alpha^{-2} |f_{\alpha} (f_{\alpha} - 1)| = c^{-2} f_t^2 \quad (366)$$

which can be written equivalently as

$$\sum_{\alpha} \alpha^{-2} f_{\alpha} (1 - f_{\alpha}) = c^{-2} f_t^2 \quad (367)$$

or as

$$f_t = \pm c \left[\sum_{\alpha} \alpha^{-2} f_{\alpha} (1 - f_{\alpha}) \right]^{1/2} \quad (368)$$

and therefore $f_t \geq 0$ but always $0 \leq f_{\alpha} \leq 1$. The condition in equations (366) or (367) relates the constants f_{α} and f_t and is required in order to insure that equations (362) and (363) are a solution to the wave equation (289). For

this special case the phase angle equations (130) and (161) give

$$\beta_{\alpha\alpha}^{uc} = -\pi/2 \quad \beta_{uu}^{ui} = \pi/2 \quad (369)$$

For a one dimensional space the solution is

$$\Psi = A \exp(j\theta_\Psi) \quad (370)$$

$$\theta_\Psi = f_x \theta_x + f_t t + f_o \quad (371)$$

subject to the condition

$$(c/x)^2 f_x (1 - f_x) = f_t^2 \quad (372)$$

or equivalently as

$$f_t = \pm c/x [f_x (1 - f_x)]^{1/2} \quad (373)$$

where f_x is a positive number satisfying $0 \leq f_x \leq 1$ while f_t can be positive or negative. This case corresponds to the propagation of the internal phase angle θ_Ψ of the wave in external time and internal space.

Case g. Ultrafast Process, Incoherent Space and Coherent Time.

For this case the wave equation (300) is valid and has a solution of the form

$$\bar{\Psi} = A \exp(j\theta_\Psi) \quad (374)$$

$$\theta_\Psi = \sum_{\alpha} g_{\alpha} \alpha + g_t \theta_t^{\Psi} + g_o \quad (375)$$

where α and θ_t^{Ψ} are variables and $t = \text{constant}$, and where g_{α} , g_t and $g_o = \text{constants}$. For this special case equations (79), (80), (209) and (210) give

$$E_{\Psi\alpha}^{ui} = g_{\alpha}^2 \quad F_{\Psi\alpha}^{ui} = 0 \quad (376)$$

$$E_{\Psi t}^{uc} = g_t (g_t - 1) \quad F_{\Psi t}^{uc} = 0 \quad (377)$$

Equations (209) and (377) give $0 \leq g_t \leq 1$. Combining equation (300) with equations (376) and (377) gives the condition required for equations (374) and (375) to be a solution of the wave equation (300) as follows

$$\begin{aligned} \sum_{\alpha} g_{\alpha}^2 &= (ct)^{-2} |g_t (g_t - 1)| \\ &= (ct)^{-2} g_t (1 - g_t) \end{aligned} \quad (378)$$

Equation (378) then gives

$$g_t = 1/2 \pm 1/2[1 - (2ct)^2 \sum_{\alpha} g_{\alpha}^2]^{1/2} \quad (379)$$

subject to

$$\sum_{\alpha} g_{\alpha}^2 \leq (2ct)^{-2} \quad (380)$$

The constants g_{α} can be positive or negative real numbers, while g_t is always a positive real number. For this special case the phase angle equations (78) and (213) give

$$\beta_{vava}^{ui} = \pi/2 \quad \beta_{uu}^{uc} = -\pi/2 \quad (381)$$

For the case of one spatial dimension the solution to equation (301) is written as

$$\bar{\psi} = A \exp(j\theta_{\psi}) \quad (382)$$

$$\theta_{\psi} = g_x x + g_t \theta_t^{\psi} + g_o \quad (383)$$

subject to the following equation that connects g_x and g_t

$$\begin{aligned} g_x^2 &= (ct)^{-2} |g_t(g_t - 1)| \\ &= (ct)^{-2} g_t(1 - g_t) \end{aligned} \quad (384)$$

with $0 \leq g_t \leq 1$. Equation (384) gives

$$g_t = 1/2 \pm 1/2[1 - (2ct)^2 g_x^2]^{1/2} \quad (385)$$

with

$$g_x \leq (2ct)^{-1} \quad (386)$$

and

$$g_x = \pm (ct)^{-1} [g_t(1 - g_t)]^{1/2} \quad (387)$$

so that $g_x \geq 0$.

Case h. Ultrafast Process, Coherent Space and Coherent Time.

A solution to equation (311) can be written as

$$\psi = A \exp(j\theta_{\psi}) \quad (388)$$

$$\theta_{\psi} = \sum_{\alpha} h_{\alpha} \theta_{\alpha} + h_t \theta_t^{\psi} + h_o \quad (389)$$

where θ_{α} and θ_t^{ψ} are variables, α and t are constants, and h_{α} , h_t and h_o are

constants. For this special case equations (126), (127), (209) and (210) give

$$E_{\psi\alpha}^{uc} = h_{\alpha}(h_{\alpha} - 1) \quad F_{\psi\alpha}^{uc} = 0 \quad (390)$$

$$F_{\psi t}^{uc} = h_t(h_t - 1) \quad F_{\psi t}^{uc} = 0 \quad (391)$$

Equations (126), (209), (390) and (391) give $0 \leq h_{\alpha} \leq 1$ and $0 \leq h_t \leq 1$ so that h_{α} and h_t are positive real numbers. Combining equation (311) with equations (390) and (391) gives

$$\sum_{\alpha} \alpha^{-2} |h_{\alpha}(h_{\alpha} - 1)| = (ct)^{-2} |h_t(h_t - 1)| \quad (392)$$

or since $0 \leq h_{\alpha} \leq 1$ and $0 \leq h_t \leq 1$ equation (392) can be rewritten as

$$\sum_{\alpha} \alpha^{-2} h_{\alpha}(1 - h_{\alpha}) = (ct)^{-2} h_t(1 - h_t) \quad (393)$$

as the condition relating the internal space parameters h_{α} and the internal time parameter h_t in order for equations (388) and (389) to be a solution to the wave equation (311). From equation (393) it follows that

$$h_t = 1/2 \pm 1/2 [1 - (2ct)^2 \sum_{\alpha} \alpha^{-2} h_{\alpha}(1 - h_{\alpha})]^{1/2} \quad (394)$$

with

$$\sum_{\alpha} \alpha^{-2} h_{\alpha}(1 - h_{\alpha}) \leq (2ct)^{-2} \quad (395)$$

For this special case the phase angle equations (130) and (213) become

$$\beta_{vava}^{uc} = -\pi/2 \quad \beta_{uu}^{uc} = -\pi/2 \quad (396)$$

For a one dimensional space the solution to equation (312) is

$$\bar{\psi} = A \exp(j\theta_{\psi}) \quad (397)$$

$$\theta_{\psi} = h_x \theta_x + h_t \theta_t^{\psi} + h_o \quad (398)$$

subject to $0 \leq h_x \leq 1$ and $0 \leq h_t \leq 1$, and equation (392) becomes

$$|h_x(h_x - 1)| = [x/(ct)]^2 |h_t(h_t - 1)| \quad (399)$$

Because $0 \leq h_x \leq 1$ and $0 \leq h_t \leq 1$ equation (399) can be written as

$$h_x(1 - h_x) = [x/(ct)]^2 h_t(1 - h_t) \quad (400)$$

which relates h_x and h_t . From equation (400) it follows that

$$h_t = 1/2 \pm 1/2 [1 - (2ct/x)^2 h_x(1 - h_x)]^{1/2} \quad (401)$$

with

$$h_x(1 - h_x) \leq [x/(2ct)]^2 \quad (402)$$

This case represents wave propagation of the internal phase θ_ψ in internal space and internal time.

Consider now an alternative analysis of the special case of an ultrafast wave propagation process in coherent space and coherent time. The solution to the wave equation (16) is written as

$$\bar{\Psi} = \Psi \exp(j\theta_\Psi) = \bar{X}_x \bar{X}_y \bar{X}_z \bar{T} \quad (403)$$

Then it follows that

$$\Psi = X_x X_y X_z T \quad (404)$$

$$\begin{aligned} \theta_\Psi &= \sum_{\alpha} \theta_{X\alpha} + \theta_T \\ &= \theta_{Xx} + \theta_{Xy} + \theta_{Xz} + \theta_T \end{aligned} \quad (405)$$

Substituting equation (403) into equation (16) yields the following complex number generalization of the standard scalar time equation and the standard one dimensional scalar homogeneous Helmholtz equations¹⁸⁻²⁴

$$d^2 \bar{T} / d\bar{t}^2 + \bar{\omega}^2 \bar{T} = 0 \quad (406)$$

$$d^2 \bar{X}_\alpha / d\bar{a}^2 + \bar{k}_\alpha^2 \bar{X}_\alpha = 0 \quad (407)$$

where $\alpha = x, y, z$ and

$$\bar{T} = T \exp(j\theta_T) \quad (408)$$

$$\bar{X}_\alpha = X_\alpha \exp(j\theta_{X\alpha}) \quad (409)$$

and where the complex number angular frequency and wave number are written as

$$\bar{\omega} = \omega \exp(j\theta_\omega) \quad (410)$$

$$\bar{k}_\alpha = k_\alpha \exp(j\theta_{k\alpha}) \quad (411)$$

It is required to solve equations (406) and (407).

The forms of the complex number second derivatives that appear in equations (406) and (407) have already appeared in equations (137) and (220). Applying equations (137) and (220) to equations (406) and (407) and dividing

through by \bar{T} and \bar{X}_α respectively gives

$$d\theta_T/d\theta_t^\Psi (1 - d\theta_T/d\theta_t^\Psi) + jd^2\theta_T/d\theta_t^{\Psi 2} = \bar{t}^2\omega^2 \quad (412)$$

$$d\theta_{X\alpha}/d\theta_\alpha (1 - d\theta_{X\alpha}/d\theta_\alpha) + jd^2\theta_{X\alpha}/d\theta_\alpha^2 = \bar{\alpha}^2 k_\alpha^2 \quad (413)$$

where t and α = constants. But for wave propagation to be possible it has been shown that $\bar{\omega}\bar{t}$ and $\bar{k}_\alpha\bar{\alpha}$ must be real numbers.¹⁷ Therefore

$$\bar{t}^2\omega^2 = t^2\omega^2 \quad \theta_\omega = -\theta_t \quad (414)$$

$$\bar{\alpha}^2 k_\alpha^2 = \alpha^2 k_\alpha^2 \quad \theta_{k\alpha} = -\theta_\alpha \quad (415)$$

Combining equations (412) through (415) gives

$$d\theta_T/d\theta_t^\Psi (1 - d\theta_T/d\theta_t^\Psi) = t^2\omega^2 \quad (416)$$

$$d^2\theta_T/d\theta_t^{\Psi 2} = 0 \quad (417)$$

$$d\theta_{X\alpha}/d\theta_\alpha (1 - d\theta_{X\alpha}/d\theta_\alpha) = \alpha^2 k_\alpha^2 \quad (418)$$

$$d^2\theta_{X\alpha}/d\theta_\alpha^2 = 0 \quad (419)$$

The solutions to equations (416) through (419) are given by

$$\theta_T = h_t \theta_t^\Psi + c_1 \quad (420)$$

$$\theta_{X\alpha} = h_\alpha \theta_\alpha + c_2 \quad (421)$$

provided that

$$h_t(1 - h_t) = t^2\omega^2 \quad (422)$$

$$h_\alpha(1 - h_\alpha) = \alpha^2 k_\alpha^2 \quad (423)$$

so that $0 \leq h_t \leq 1$ and $0 \leq h_\alpha \leq 1$. Equivalently, equations (422) and (423) can be written as

$$h_t = 1/2 \pm 1/2(1 - 4t^2\omega^2)^{1/2} \quad (424)$$

$$h_\alpha = 1/2 \pm 1/2(1 - 4\alpha^2 k_\alpha^2)^{1/2} \quad (425)$$

subject to

$$\omega \leq (2t)^{-1} \quad k_\alpha \leq (2\alpha)^{-1} \quad (426)$$

Equations (424) and (425) are equivalent to equation (394), and equation (426) is equivalent to equation (395), as can be seen by noting that equations (422) through (425) are equivalent to equation (394) because it follows from equation (423)

$$\sum_\alpha \alpha^{-2} h_\alpha (1 - h_\alpha) = \sum_\alpha k_\alpha^2 = k^2 = \omega^2/c^2 \quad (427)$$

Combining equations (405), (420) and (421) gives

$$\theta_\psi = \sum_\alpha h_\alpha \theta_\alpha + h_t \theta_t^\psi + h_o \quad (428)$$

which is just equation (389). Equations (412) and (413) or equivalently equations (416) through (419) describe internal phase wave propagation that occurs at a fixed point (α, t) in spacetime.

5. CONCLUSION. The wave equation is somewhat more complex than it is generally thought to be. In the presence of an external field such as gravitation or electromagnetism, the wave amplitude and the space and time coordinates can be taken to be complex numbers in an internal space. For this case the wave equation has eight possible limiting forms according to whether the wave amplitude changes slowly or in an ultrafast manner and according to whether the change in the space coordinates is coherent or incoherent and the change in the time coordinates is coherent or incoherent. For a slow wave propagation process the wave amplitude changes in magnitude whereas for the case of an ultrafast wave propagation process the wave function changes by a rotation in internal space. Similarly, incoherent space and time changes occur through a variation of the magnitudes of the space and time coordinates, whereas coherent space and time changes occur as rotations in an internal space. The eight possible forms of the wave equation have unique solutions with their own special characteristics and suggest that new and as yet unobserved phenomena can occur for wave propagation in matter in the presence of strong external fields such as may be realized in stellar compact objects.

ACKNOWLEDGEMENT

This work would not have been completed without the kind assistance of Elizabeth K. Klein who typed and edited this paper.

REFERENCES

1. Murnane, M. M., Kapteyn, H. C., Rosen, M. D. and Falcone, R. W., "Ultrafast X-ray Pulses from Laser-Produced Plasmas," Science, Vol. 251, p. 531, 1 February 1991.
2. Nibbering, E. T. J., Wiersma, D. A. and Duppen, K., "Femtosecond Non-Markovian Optical Dynamics in Solution," Phys. Rev. Lett., Vol. 66, p. 2464, 13 May 1991.

3. Grinberg, A. A. and Luryi, S., "Nonstationary Quasiperiodic Energy Distribution of an Electron Gas upon Ultrafast Thermal Excitation," Phys. Rev. Lett., Vol. 65, p. 1251, 3 September 1990.
4. Binder, R., Koch, S. W., Lindberg, M. and Peyghambarian, N., "Ultrafast Adiabatic Following in Semiconductors," Phys. Rev. Lett., Vol. 65, p. 899, 13 August 1990.
5. Fork, R. L., Avramopoulos, H. and Valdmanis, J. A., "Ultrashort Light Pulses," American Scientist, Vol. 78, p. 216, May-June 1990.
6. Zewail, A. H., "The Birth of Molecules," Scientific American, p. 76, December 1990.
7. Bokor, J., "Ultrafast Dynamics at Semiconductor and Metal Surfaces," Science, Vol. 246, p. 1130, 1 December 1989.
8. Murnane, M. M., Kapeyn, H. C. and Falcone, R. W., "High-Density Plasmas Produced by Ultrafast Laser Pulses," Phys. Rev. Lett., Vol. 62, p. 155, 9 January 1989.
9. Zewail, A. H., "Laser Femtochemistry," Science, Vol. 242, p. 1645, 23 December 1988.
10. Hicks, J. M., Urbach, L. E., Plummer, E. W. and Dai, H. L., "Can Pulsed Excitation of Surfaces Be Described by a Thermal Model?," Phys. Rev. Lett., Vol. 61, p. 2588, 28 November 1988.
11. Peters, K. S. and Snyder, G. J., "Time-Resolved Photoacoustic Calorimetry: Probing the Energetics and Dynamics of Fast Chemical and Biochemical Reactions," Science, Vol. 241, p. 1053, 26 August 1988.
12. Yablonovitch, E., "Energy Conservation in the Picosecond and Subpicosecond Photoelectric Effect," Phys. Rev. Lett., Vol. 60, p. 795, 29 February 1988.
13. Davis, W. C., "The Detonation of Explosives," Scientific American, p. 106, May 1987.
14. Trimble, V., "1987A: The Greatest Supernova Since Kepler," Revs. Mod. Phys., Vol. 60, p. 859, October 1988.
15. Woosley, S. E. and Phillips, M. M., "Supernova 1987A," Science, Vol. 240, p. 750, 6 May 1988.
16. Fowler, W. A., "Experimental and Theoretical Nuclear Astrophysics: The Quest for the Origin of the Elements," Rev. Mod. Phys., Vol. 56, p. 149, April 1984.
17. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.
18. Page, L., Introduction to Theoretical Physics, Van Nostrand, New York, 1952.

19. Morse, P. M. and Feshbach, H., Methods of Mathematical Physics, Vols. 1 and 2, McGraw-Hill, New York, 1953.
20. Bateman, H., Partial Differential Equations of Mathematical Physics, Dover, New York, 1944.
21. Jeffreys, H. and Swirles, B., Methods of Mathematical Physics, Cambridge, New York, 1956.
22. Page, C. H., Physical Mathematics, Van Nostrand, New York, 1955.
23. Coulson, C. A., Waves, Oliver and Boyd, London, 1952.
24. Sneddon, I. N., Fourier Transforms, McGraw-Hill, New York, 1951.
25. Weiss, R. A., "Electromagnetism and Gravity," Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 91-1, June 19-22, p. 265, 1990.

CLEAN FISSION NUCLEAR REACTORS

Richard A. Weiss

U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. This paper develops the basic physics concepts necessary for the design of clean fission nuclear reactors. The concepts are based on the creation of a broken symmetry state of the atomic number and the atomic mass number when nuclei are placed in an electromagnetic field. In an electromagnetic field the atomic number and atomic mass number of a nucleus must be represented by special types of complex numbers in an internal space. The nuclear radius, binding energy and mass must also be represented by complex numbers, and therefore a liquid drop model type of nuclear mass formula is developed that incorporates complex number values of the atomic number and atomic mass number. The Bohr-Wheeler condition for spontaneous and thermal neutron induced nuclear fission is evaluated for nuclei in the presence of an electromagnetic field. The electromagnetic field introduces the internal phase angles of the atomic number and atomic mass number into the fission condition in such a way that nuclear fission with thermal neutrons can occur in nuclei lighter than the actinides. It is found that the wavelength of the electromagnetic waves required for the catalysis of clean fission nuclear reactions with thermal neutrons is in the γ ray region, and in particular that the energy of the γ rays corresponds to the giant dipole resonance frequency of the subactinide element that is chosen as nuclear fuel. This suggests that nuclei in the presence of γ rays can be fissioned by thermal neutrons even for values of Z and A sufficiently small as to prohibit fission under zero field conditions. A formula is presented that allows the calculation of the energy and power density of the γ rays required to catalyze the sustained nuclear fission of elements lighter than the actinides by thermal neutrons. The γ ray catalyzed clean fission nuclear reactors will produce only low level nuclear waste products or possibly no radioactive wastes at all. Design concepts for a γ ray catalyzed thermal neutron induced clean fission nuclear reactor are presented which will deliver clean power with little or no radioactive waste products. In fact, the nuclear wastes of present day uranium and plutonium reactors can be used as fuel elements in clean fission nuclear reactors, and in this way stored waste radionuclides can be eliminated in a useful way to generate power.

1. INTRODUCTION. The generation of power has always been at high cost to living organisms. For millennia and still today beasts of burden were used to generate power for agricultural purposes, transportation and mining.¹ With the advent of the industrial revolution came the engines and machines which required wood, coal and petroleum products as fuel. The combustion of millions of tons of coal and gallons of gasoline pollutes the atmosphere with carbon dioxide and sulfur dioxide among many other chemical compounds some of which are carcinogens.²⁻⁸ These pollutants enter the atmosphere and cause damage to forests by producing acid rain, and damage to people by an increased incidence of lung cancer, emphysema and other diseases.²⁻⁸ Nuclear fission reactors, once thought to be the power source of the future, are now known to be a potentially

dangerous way of generating power because they are susceptible to accidents and because they produce radioactive waste products whose safe disposal still challenges engineers and scientists a half century after the first nuclear reactor was developed.⁹⁻¹⁵ After the Three Mile Island and Chernobyl nuclear accidents the lack of public confidence in fission reactor power plants has brought the construction of new power plants to a standstill.

Alternative energy sources are now one of the major research endeavors of science and engineering.¹⁶⁻²² Wind power has been utilized on a small scale in a few areas but it cannot power a city like Chicago.²³⁻²⁵ Solar power can be used effectively to heat single family dwellings in geographic areas which have generally clear skies, but it is of limited use on a large scale and cannot be used to power New York City.²⁶⁻²⁹ Geothermal energy is clearly of use only in limited geographic regions and has produced no significant contribution to the world's power generation capacity and it cannot be used to power a city like Tokyo.³⁰⁻³² Nuclear fusion power still remains a dream after nearly a half century of research and billions of dollars of investment.³³⁻⁴⁵ All of these energy sources are impractical alternatives to nuclear fission power. If the safety and nuclear waste problems can be solved it appears that nuclear fission reactors still represent the best hope for future power development.⁴⁶⁻⁵⁵ But new concepts and designs for clean and environmentally safe nuclear reactors are required. This paper presents the mathematical and physical theory for the development of γ ray catalyzed clean fission nuclear reactors which have little or no dangerous radionuclide waste products, and which are operationally safe. Design concepts for clean fission nuclear reactor cores are presented.

The clean fission nuclear processes that are described in this paper are based on the concept of the broken symmetry of quantum numbers and space and time coordinates.⁵⁶ It has been suggested that space and time coordinates have internal phase angles and are written as complex numbers as follows⁵⁶

$$\bar{x} = x \exp(j\theta_x) \quad \bar{y} = y \exp(j\theta_y) \quad \bar{z} = z \exp(j\theta_z) \quad (1)$$

$$\bar{t} = t \exp(j\theta_t) \quad (2)$$

where the phase angles describe an internal space. For spherical polar coordinates the complex number space coordinates are written as⁵⁶

$$\bar{r} = r \exp(j\theta_r) \quad \bar{\phi} = \phi \exp(j\theta_\phi) \quad \bar{\psi} = \psi \exp(j\theta_\psi) \quad (3)$$

Corresponding to the complex number azimuthal angle is the following complex magnetic quantum number⁵⁶

$$\bar{M} = M \exp(j\theta_M) = m \cos \theta_\phi \exp(-j\theta_\phi) \quad (4)$$

where $m = 0, \pm 1, \pm 2, \pm 3, \dots$ is the standard magnetic quantum number. Therefore⁵⁶

$$M = m \cos \theta_\phi \quad \theta_M = -\theta_\phi \quad (5)$$

It is often convenient to define the following positive magnetic quantum number⁵⁶

$$\bar{M}' = M' \exp(j\theta_M') = |m| \cos \theta_\phi \exp(-j\theta_\phi) \quad (6)$$

$$M' = |m| \cos \theta_\phi \quad \theta_M' = -\theta_\phi \quad (7)$$

For $\theta_\phi = 0$ these quantities reduce to the standard concepts.

Under the action of a time varying external field, such as an electromagnetic wave, or simply a spontaneous decay, the magnetic quantum number has the following time variation

$$d\bar{M}/dt = \bar{F}_M \cos \beta_{tt} \exp[j(\theta_M - \theta_t - \beta_{tt})] \quad (8)$$

where

$$\begin{aligned} \bar{F}_M &= dm/dt \cos \theta_M - m \sin \theta_M d\theta_M/dt + jm \cos \theta_M d\theta_M/dt \\ &= dm/dt \cos \theta_M + md\theta_M/dt \exp[j(\pi/2 + \theta_M)] \\ &= dm/dt \cos \theta_\phi - m \sin \theta_\phi d\theta_\phi/dt - jm \cos \theta_\phi d\theta_\phi/dt \\ &= dm/dt \cos \theta_\phi - md\theta_\phi/dt \exp[j(\pi/2 - \theta_\phi)] \end{aligned} \quad (9)$$

and where

$$\tan \beta_{tt} = t \partial \theta_t / \partial t \quad (10)$$

It is assumed that $\theta_M = \theta_M(m, H)$ and $\theta_\phi = \theta_\phi(m, H)$ where H = applied magnetic field strength. The application of the chain rule gives the following derivatives

$$d\theta_M/dt = \partial \theta_M / \partial m dm/dt + \partial \theta_M / \partial H dH/dt \quad (11)$$

$$d\theta_\phi/dt = \partial \theta_\phi / \partial m dm/dt + \partial \theta_\phi / \partial H dH/dt \quad (12)$$

Then equation (8) can be written as

$$d\bar{M}/dt = (\bar{B}_M dm/dt + \bar{C}_M dH/dt) \cos \beta_{tt} \exp[j(\theta_M - \theta_t - \beta_{tt})] \quad (13)$$

where

$$\begin{aligned} \bar{B}_M &= \cos \theta_M + m \partial \theta_M / \partial m \exp[j(\pi/2 + \theta_M)] \\ &= \cos \theta_\phi - m \partial \theta_\phi / \partial m \exp[j(\pi/2 - \theta_\phi)] \end{aligned} \quad (14)$$

$$\begin{aligned} \bar{C}_M &= m \partial \theta_M / \partial H \exp[j(\pi/2 + \theta_M)] \\ &= -m \partial \theta_\phi / \partial H \exp[j(\pi/2 - \theta_\phi)] \end{aligned} \quad (15)$$

Two special cases are of interest.

Case a. Constant Magnetic Field.

For this case equation (13) becomes

$$(d\bar{M}/d\bar{t})_H = \bar{B}_M \cos \beta_{tt} dm/dt \exp[j(\theta_M - \theta_t - \beta_{tt})] \quad (16)$$

which corresponds to spontaneous decay for a fixed magnetic field.

Case b. Constant m.

From equation (13) and (15) it follows that for this case

$$\begin{aligned} (d\bar{M}/d\bar{t})_m &= \bar{C}_M dH/dt \cos \beta_{tt} \exp[j(\theta_M - \theta_t - \beta_{tt})] \\ &= m \partial \theta_M / \partial H dH/dt \cos \beta_{tt} \exp[j(\pi/2 + 2\theta_M - \theta_t - \beta_{tt})] \\ &= -m \partial \theta_\phi / \partial H dH/dt \cos \beta_{tt} \exp[j(\pi/2 - 2\theta_\phi - \theta_t - \beta_{tt})] \\ &= \bar{M} \sec \theta_M \partial \theta_M / \partial H dH/dt \cos \beta_{tt} \exp[j(\pi/2 + \theta_M - \theta_t - \beta_{tt})] \end{aligned} \quad (17)$$

which corresponds to a decay associated with a change in the internal phase angle of the azimuthal angle that is induced by a time varying magnetic field.

A more general expression for the time derivative of the complex magnetic quantum number can be written as

$$d\bar{M}/d\bar{t} = \sec \beta_{MM} \cos \beta_{tt} dM/dt \exp(j\phi_{Mt}) \quad (18)$$

$$= \csc \beta_{MM} \cos \beta_{tt} M d\theta_M/dt \exp(j\phi_{Mt}) \quad (19)$$

$$= \sec \beta_{MM} \sin \beta_{tt} t^{-1} dM/d\theta_t \exp(j\phi_{Mt}) \quad (20)$$

$$= \csc \beta_{MM} \sin \beta_{tt} M/t d\theta_M/d\theta_t \exp(j\phi_{Mt}) \quad (21)$$

where

$$\tan \beta_{MM} = M \partial \theta_M / \partial M = -M \partial \theta_\phi / \partial M \quad (22)$$

$$\phi_{Mt} = \theta_M + \beta_{MM} - \theta_t - \beta_{tt} \quad (23)$$

From equation (5) it follows that the time rate of change of the magnitude of the magnetic quantum number is given by

$$dM/dt = dm/dt \cos \theta_\phi - m \sin \theta_\phi d\theta_\phi/dt \quad (24)$$

Combining equations (12) and (24) gives

$$\begin{aligned} dM/dt &= dm/dt [\cos \theta_\phi - m \sin \theta_\phi \partial \theta_\phi / \partial m] - m \sin \theta_\phi \partial \theta_\phi / \partial H dH/dt \\ &= dm/dt [\cos \theta_M - m \sin \theta_M \partial \theta_M / \partial m] - m \sin \theta_M \partial \theta_M / \partial H dH/dt \end{aligned} \quad (25)$$

$$\cos^2 \theta_M = 1/2[1 + (1 - 4f^2)^{1/2}] \quad (38)$$

$$f = m^{-1}(m_1 \sin \theta_{M1} \cos \theta_{M1} + m_2 \sin \theta_{M2} \cos \theta_{M2}) \quad (39)$$

where m is given by equation (34).

This brief summary of the theory of broken symmetry space and time coordinates and broken symmetry magnetic quantum numbers is sufficient to begin a study of the effects of broken symmetry on the quantum numbers of atomic nuclei. It is suggested in this paper that in the presence of a γ ray electromagnetic field the atomic number Z and the atomic mass number A are integer quantum numbers, analogous to $|m|$ of equation (7), which have complex number generalizations \bar{z} and \bar{a} respectively, analogous to \bar{M}' of equation (7), which determine the binding energy, symmetry energy, Coulomb energy, nuclear surface energy, giant resonance frequency, incompressibility, fissility and many other nuclear properties. It is also suggested that the number of nuclei located in an electromagnetic field or gravitational field has an associated internal phase angle, which allows the possibility of coherent radioactive decays in which the integer number of nuclei is fixed but the internal phase angle of the number of nuclei is changing. The theory of nuclear fission will be affected by the concept that the atomic number and atomic mass number must be taken to be complex numbers in an internal space when nuclei are located in the presence of an electromagnetic field. It is well known that some of the actinide nuclei elements either fission spontaneously or undergo induced fission by thermal neutrons, but present day theory suggests that spontaneous or thermal neutron induced fission does not occur for nuclei lighter than the actinides. The main thrust of this paper is the development of a theory of γ ray catalyzed spontaneous and thermal neutron induced fission reactions of nuclei lighter than the actinides but heavier than ^{56}Fe . The theory suggests that thermal neutron induced clean fission of subactinide elements can be accomplished by placing the potential nuclear fuel in a γ ray field whose energy (frequency) is determined to correspond to the giant dipole resonance frequency of the subactinide fuel element, and whose power density is adjusted until the internal phase angle of the atomic number attains a critical value at which point the fuel nuclei are in a state for which clean nuclear fission is possible. Because the fuel element to be fissioned has a relatively low atomic mass number, such as ^{63}Cu for example, the fission products will also be relatively light elements which are in or not far removed from the valley of beta stability and therefore are either not radioactive or are low level beta, alpha, and neutron emitters. These lighter than actinide nuclei cannot be used as fuel in conventional reactors.

An outline of this paper is as follows: Section 2 introduces the concept of complex atomic numbers and complex atomic mass numbers and investigates their addition laws and time dependence, Section 3 considers the radioactive decay of nuclei located in an electromagnetic field, Section 4 develops a liquid drop model type of nuclear mass formula for nuclei in an electromagnetic field, Section 5 investigates the effect of an electromagnetic field on the spontaneous and thermal neutron induced fission condition for nuclei and demonstrates the possibility of clean fission in lighter than actinide nuclei, Section 6 studies the final state energy conditions for the clean fission of nuclei lighter than the actinides but heavier than ^{56}Fe , Section 7 presents numerical calculations

Combining equations (11), (18), (19) and (25) gives

$$\begin{aligned} d\bar{M}/dt &= J_M \cos \beta_{tt} \sec \beta_{MM} \cos \theta_M \exp(j\phi_{Mt}) \\ &= I_M \cos \beta_{tt} \csc \beta_{MM} \cos \theta_M \exp(j\phi_{Mt}) \\ &= (I_M^2 + J_M^2)^{1/2} \cos \beta_{tt} \cos \theta_M \exp(j\phi_{Mt}) \end{aligned} \quad (26)$$

where now

$$\tan \beta_{MM} = I_M/J_M \quad (27)$$

with

$$I_M = \tan \alpha_{mm}^H dm/dt + m/H \tan \alpha_{HH}^m dH/dt \quad (28)$$

$$J_M = (1 - \tan \theta_M \tan \alpha_{mm}^H) dm/dt - m/H \tan \theta_M \tan \alpha_{HH}^m dH/dt \quad (29)$$

where

$$\tan \alpha_{mm}^H = m \partial \theta_M / \partial m \quad (30)$$

$$\tan \alpha_{HH}^m = H \partial \theta_M / \partial H \quad (31)$$

Equation (26) is equivalent to equation (13).

Consider two complex magnetic quantum numbers which can be written as

$$\bar{M}_1 = M_1 \exp(j\theta_{M1}) = m_1 \cos \theta_{M1} \exp(j\theta_{M1}) \quad (32)$$

$$\bar{M}_2 = M_2 \exp(j\theta_{M2}) = m_2 \cos \theta_{M2} \exp(j\theta_{M2}) \quad (33)$$

where the following integer addition law is always true

$$m = m_1 + m_2 \quad (34)$$

where m, m_1 and $m_2 = 0, \pm 1, \pm 2, \pm 3, \dots$. Then the addition law for the complex magnetic quantum numbers is written as

$$\bar{M} + W = \bar{M}_1 + \bar{M}_2 \quad (35)$$

where

$$\bar{M} = M \exp(j\theta_M) = m \cos \theta_M \exp(j\theta_M) \quad (36)$$

where the unknown quantities W and θ_M are given by⁵⁶

$$W = -m/2[1 + (1 - 4f^2)^{1/2}] + m_1 \cos^2 \theta_{M1} + m_2 \cos^2 \theta_{M2} \quad (37)$$

of the γ ray photon energy and flux density that are required to catalyze the clean fission of subactinide nuclei and gives examples of clean fission nuclear reactions, finally Section 8 presents design concepts for a γ ray catalyzed thermal neutron induced clean fission nuclear reactor.

2. ATOMIC NUCLEI WITH BROKEN INTERNAL SYMMETRIES. This section considers the broken internal symmetries associated with the atomic number, neutron number and atomic mass number of nuclei that are located in an electromagnetic or gravitational field.

A. Complex Atomic Number, Neutron Number and Atomic Mass Number.

It is assumed that the integer quantum numbers Z , N and A are analogous to the magnetic quantum number $|m|$ and occur in a solution to an azimuthal portion of a wave equation in internal space.⁵⁶ Then by an argument similar to that for the complex number magnetic quantum number it follows that the atomic number, neutron number and atomic mass number are complex numbers in an internal space and can be written in a form similar to that of the complex magnetic quantum number \bar{M}' in equation (6) as follows⁵⁶

$$\bar{z} = z \exp(j\theta_z) = Z \cos \theta_z \exp(j\theta_z) \quad (40)$$

$$\bar{n} = n \exp(j\theta_n) = N \cos \theta_n \exp(j\theta_n) \quad (41)$$

$$\bar{a} = a \exp(j\theta_a) = A \cos \theta_a \exp(j\theta_a) \quad (42)$$

where the complex number azimuthal angles corresponding to the quantum numbers \bar{z} , \bar{n} and \bar{a} are written, in analogy to the azimuthal angle of real space given in equation (3), as follows

$$\bar{\phi}_z = \phi_z \exp(j\theta_{\phi z}) \quad \bar{\phi}_n = \phi_n \exp(j\theta_{\phi n}) \quad \bar{\phi}_a = \phi_a \exp(j\theta_{\phi a}) \quad (43)$$

where $\theta_{\phi z}$, $\theta_{\phi n}$ and $\theta_{\phi a}$ are the internal phase angles of the complex number azimuthal angles in the internal space of nucleons.⁵⁶ The wave equations in internal space are written, in analogy to the complex number azimuthal equation of Reference 56, as

$$d^2 \bar{\phi}_z / d\bar{\phi}_z^2 + \bar{z}^2 \bar{\phi}_z = 0 \quad (44)$$

$$d^2 \bar{\phi}_n / d\bar{\phi}_n^2 + \bar{n}^2 \bar{\phi}_n = 0 \quad (45)$$

$$d^2 \bar{\phi}_a / d\bar{\phi}_a^2 + \bar{a}^2 \bar{\phi}_a = 0 \quad (46)$$

The nuclear wave function has a form which is analogous to the azimuthal wave function of the central field problem with a broken internal symmetry of the azimuthal angle and is written as

$$\bar{\phi} = \bar{\phi}_z \bar{\phi}_n \bar{\phi}_a = \bar{C} \exp[i(\bar{z}\bar{\phi}_z + \bar{n}\bar{\phi}_n + \bar{a}\bar{\phi}_a)] \quad (47A)$$

whereas the azimuthal wave function of the central field problem with broken internal symmetries is written as⁵⁶

$$\phi = \bar{C} \exp(i\bar{M}\bar{\phi}) \quad (47B)$$

Strictly speaking \bar{z} , \bar{n} and \bar{a} correspond to \bar{M}' of equation (6). The measured values of the angles that appear in equation (47A) are given by

$$\phi_{zm} = \phi_z \cos \theta_{\phi z} \quad \phi_{nm} = \phi_n \cos \theta_{\phi n} \quad \phi_{am} = \phi_a \cos \theta_{\phi a} \quad (48)$$

If the wave function is to remain unchanged under the transformations

$$\phi_{zm} \rightarrow \phi_{zm} + 2\pi \quad \phi_{nm} \rightarrow \phi_{nm} + 2\pi \quad \phi_{am} \rightarrow \phi_{am} + 2\pi \quad (49)$$

then the argument in equation (47A) must be a real number

$$\bar{z}\bar{\phi}_z = z\phi_z \quad \theta_z + \theta_{\phi z} = 0 \quad (50)$$

$$\bar{n}\bar{\phi}_n = n\phi_n \quad \theta_n + \theta_{\phi n} = 0 \quad (51)$$

$$\bar{a}\bar{\phi}_a = a\phi_a \quad \theta_a + \theta_{\phi a} = 0 \quad (52)$$

By writing the complex quantum numbers \bar{z} , \bar{n} and \bar{a} as

$$\bar{z} = z_R + jz_I = z(\cos \theta_z + j \sin \theta_z) \quad (53)$$

$$\bar{n} = n_R + jn_I = n(\cos \theta_n + j \sin \theta_n) \quad (54)$$

$$\bar{a} = a_R + ja_I = a(\cos \theta_a + j \sin \theta_a) \quad (55)$$

and the complex number azimuthal angles associated with \bar{z} , \bar{n} and \bar{a} as

$$\bar{\phi}_z = \phi_{zR} + j\phi_{zI} = \phi_z(\cos \theta_{\phi z} + j \sin \theta_{\phi z}) \quad (56)$$

$$\bar{\phi}_n = \phi_{nR} + j\phi_{nI} = \phi_n(\cos \theta_{\phi n} + j \sin \theta_{\phi n}) \quad (57)$$

$$\bar{\phi}_a = \phi_{aR} + j\phi_{aI} = \phi_a(\cos \theta_{\phi a} + j \sin \theta_{\phi a}) \quad (58)$$

it follows that

$$\bar{z}\bar{\phi}_z = z_R\phi_{zR} - z_I\phi_{zI} + j(z_I\phi_{zR} + z_R\phi_{zI}) \quad (59)$$

$$\bar{n}\bar{\phi}_n = n_R\phi_{nR} - n_I\phi_{nI} + j(n_I\phi_{nR} + n_R\phi_{nI}) \quad (60)$$

$$\bar{a}\bar{\phi}_a = a_R\phi_{aR} - a_I\phi_{aI} + j(a_I\phi_{aR} + a_R\phi_{aI}) \quad (61)$$

If the imaginary parts of equations (59) through (61) are zero, as they must be if equations (50) through (52) are valid, then

$$\phi_{zI} = -z_I \phi_{zR}/z_R \quad \phi_{nI} = -n_I \phi_{nR}/n_R \quad \phi_{aI} = -a_I \phi_{aR}/a_R \quad (62)$$

Substituting equation (62) into equations (59) through (61) gives

$$\bar{z}\bar{\phi}_z = (z_R^2 + z_I^2)\phi_{zR}/z_R = z^2\phi_{zR}/z_R = z^2\phi_{zm}/z_R \quad (63)$$

$$\bar{n}\bar{\phi}_n = (n_R^2 + n_I^2)\phi_{nR}/n_R = n^2\phi_{nR}/n_R = n^2\phi_{nm}/n_R \quad (64)$$

$$\bar{a}\bar{\phi}_a = (a_R^2 + a_I^2)\phi_{aR}/a_R = a^2\phi_{aR}/a_R = a^2\phi_{am}/a_R \quad (65)$$

Therefore the argument in equation (47A) is linear in ϕ_{zm} , ϕ_{nm} and ϕ_{am} .

If the wave function in equation (47A) is to remain unchanged under the transformations in equation (49) it follows from equations (63) through (65) that

$$z^2/z_R = Z \quad n^2/n_R = N \quad a^2/a_R = A \quad (66)$$

where Z , N and A are integers. In analogy to equations (5) and (7) for the magnetic quantum numbers, the combination of equations (66) with equations (53) through (55) gives

$$z = Z \cos \theta_z \quad n = N \cos \theta_n \quad a = A \cos \theta_a \quad (67)$$

$$\theta_z = -\theta_{\phi z} \quad \theta_n = -\theta_{\phi n} \quad \theta_a = -\theta_{\phi a} \quad (68)$$

and therefore in analogy to equation (6)

$$\bar{z} = Z \cos \theta_z \exp(j\theta_z) = Z \cos \theta_{\phi z} \exp(-j\theta_{\phi z}) \quad (69)$$

$$\bar{n} = N \cos \theta_n \exp(j\theta_n) = N \cos \theta_{\phi n} \exp(-j\theta_{\phi n}) \quad (70)$$

$$\bar{a} = A \cos \theta_a \exp(j\theta_a) = A \cos \theta_{\phi a} \exp(-j\theta_{\phi a}) \quad (71)$$

Sometimes it is convenient to consider the terms

$$\bar{l}_z = \cos \theta_z \exp(j\theta_z) = \cos \theta_{\phi z} \exp(-j\theta_{\phi z}) \quad (72)$$

$$\bar{l}_n = \cos \theta_n \exp(j\theta_n) = \cos \theta_{\phi n} \exp(-j\theta_{\phi n}) \quad (73)$$

$$\bar{l}_a = \cos \theta_a \exp(j\theta_a) = \cos \theta_{\phi a} \exp(-j\theta_{\phi a}) \quad (74)$$

Then equations (69) through (71) become

$$\bar{z} = Z\bar{l}_z \quad \bar{n} = N\bar{l}_n \quad \bar{a} = A\bar{l}_a \quad (75)$$

The real and imaginary parts of equations (69) through (71) are written as

$$z_R = Z \cos^2 \theta_z = Z \cos^2 \theta_{\phi z} \quad (76)$$

$$n_R = N \cos^2 \theta_n = N \cos^2 \theta_{\phi n} \quad (77)$$

$$a_R = A \cos^2 \theta_a = A \cos^2 \theta_{\phi a} \quad (78)$$

$$z_I = Z \cos \theta_z \sin \theta_z = -Z \cos \theta_{\phi z} \sin \theta_{\phi z} \quad (79)$$

$$n_I = N \cos \theta_n \sin \theta_n = -N \cos \theta_{\phi n} \sin \theta_{\phi n} \quad (80)$$

$$a_I = A \cos \theta_a \sin \theta_a = -A \cos \theta_{\phi a} \sin \theta_{\phi a} \quad (81)$$

All of these results follow from the periodicity of the wave function in the variables ϕ_{zm} , ϕ_{nm} and ϕ_{am} .

The law of addition for the complex atomic number, complex neutron number and complex atomic mass number must be analogous to the form of the addition law for complex magnetic quantum numbers given in equation (35), so that

$$W + \bar{a} = \bar{z} + \bar{n} \quad (82)$$

or

$$W + A\bar{I}_a = Z\bar{I}_z + N\bar{I}_n \quad (83)$$

subject to the integer relation of the universal law of baryon number conservation

$$A = Z + N \quad (84)$$

In equation (82) the known quantities are Z , θ_z , N and θ_n . The value of A is immediately obtained from equation (84) while the quantities W and θ_a are obtained by noting that equation (82) can be rewritten as

$$W + A \cos^2 \theta_a = Z \cos^2 \theta_z + N \cos^2 \theta_n \quad (85)$$

$$A \cos \theta_a \sin \theta_a = Z \cos \theta_z \sin \theta_z + N \cos \theta_n \sin \theta_n \quad (86)$$

From equations (85) and (86) it follows that

$$W = -A/2[1 + (1 - 4f^2)^{1/2}] + Z \cos^2 \theta_z + N \cos^2 \theta_n \quad (87)$$

$$\cos^2 \theta_a = 1/2[1 + (1 - 4f^2)^{1/2}] \quad (88)$$

$$f = A^{-1}(Z \cos \theta_z \sin \theta_z + N \cos \theta_n \sin \theta_n) \quad (89)$$

which have the same form as equations (37) through (39) for complex magnetic quantum numbers. For small internal phase angles $\theta_z \sim 0$ and $\theta_n \sim 0$, it follows that $f \sim 0$, $W \sim 0$ and $\theta_a \sim 0$. In this case the function W can be neglected to a first approximation in nuclear physics calculations as, for example, in the determination of the valley of beta stability which is done in Section 4. If the approximation $\theta_z \sim \theta_n$ is assumed then

$$f \sim \cos \theta_z \sin \theta_z \quad (89A)$$

For small values of $\theta_z \sim \theta_n$ equation (89A) becomes $f \sim \theta_z$, and equation (88) gives $\theta_a \sim \theta_z$ and therefore for small internal phase angles the following approximation is valid

$$\theta_z \sim \theta_n \sim \theta_a \quad (89B)$$

Equation (88) shows that $\theta_a = \theta_a(\theta_z, \theta_n, Z, N)$ so that θ_a is not an independent variable. For simplicity it will sometimes be assumed that $\theta_z = \theta_z(Z, H)$, $\theta_n = \theta_n(N, H)$ and $\theta_a = \theta_a(Z, N, H) \sim \theta_a(A, H)$, where H = strength of an externally applied field such as a magnetic field component of an applied electromagnetic field.

B. Time Variation of the Complex Atomic Number, Neutron Number and Atomic Mass Number.

This subsection considers the time variation of \bar{z} , \bar{n} and \bar{a} for two physical cases: the first is associated with changes of Z , N and A by radioactive transmutations in a constant magnetic field, and the second case is associated with the time variation of the magnetic field for fixed values of Z , N and A . Equations (40) through (42) give

$$d\bar{z}/dt = \cos \beta_{tt} (dz/dt + jz d\theta_z/dt) \exp[j(\theta_z - \theta_t - \beta_{tt})] \quad (90)$$

$$d\bar{n}/dt = \cos \beta_{tt} (dn/dt + jn d\theta_n/dt) \exp[j(\theta_n - \theta_t - \beta_{tt})] \quad (91)$$

$$d\bar{a}/dt = \cos \beta_{tt} (da/dt + jad\theta_a/dt) \exp[j(\theta_a - \theta_t - \beta_{tt})] \quad (92)$$

where β_{tt} is given by equation (10). It is assumed that θ_z , θ_n and θ_a are functions of an applied electromagnetic field so that for simplicity

$$\theta_z = \theta_z(Z, H) \quad \theta_n = \theta_n(N, H) \quad \theta_a = \theta_a(A, H) \quad (93)$$

where H = magnitude of a static magnetic field or the amplitude of the magnetic field component of an electromagnetic field. Actually the precise value of θ_a is given by equation (88). The magnetic field H is chosen as a descriptive value of an electromagnetic field because in Section 7 a comparison is made of the effects of a static magnetic field and an electromagnetic wave field on the clean fission process in nuclei, and the field strength H is a common variable of description for the two fields and can serve as a means of comparison.

For the case when Z , N , A and H are all time dependent the application of the chain rule for derivatives gives

$$d\theta_z/dt = \partial\theta_z/\partial Z dZ/dt + \partial\theta_z/\partial H dH/dt \quad (94)$$

$$d\theta_n/dt = \partial\theta_n/\partial N dN/dt + \partial\theta_n/\partial H dH/dt \quad (95)$$

$$d\theta_a/dt = \partial\theta_a/\partial A dA/dt + \partial\theta_a/\partial H dH/dt \quad (96)$$

Combining equations (67) with equations (90) through (96) gives

$$d\bar{z}/d\bar{t} = \cos \beta_{tt} (\bar{B}_z dZ/dt + \bar{C}_z dH/dt) \exp[j(\theta_z - \theta_t - \beta_{tt})] \quad (97)$$

$$d\bar{n}/d\bar{t} = \cos \beta_{tt} (\bar{B}_n dN/dt + \bar{C}_n dH/dt) \exp[j(\theta_n - \theta_t - \beta_{tt})] \quad (98)$$

$$d\bar{a}/d\bar{t} = \cos \beta_{tt} (\bar{B}_a dA/dt + \bar{C}_a dH/dt) \exp[j(\theta_a - \theta_t - \beta_{tt})] \quad (99)$$

where

$$\bar{B}_z = \cos \theta_z + Z \partial\theta_z/\partial Z \exp[j(\pi/2 + \theta_z)] \quad (100)$$

$$\bar{B}_n = \cos \theta_n + N \partial\theta_n/\partial N \exp[j(\pi/2 + \theta_n)] \quad (101)$$

$$\bar{B}_a = \cos \theta_a + A \partial\theta_a/\partial A \exp[j(\pi/2 + \theta_a)] \quad (102)$$

$$\bar{C}_z = Z \partial\theta_z/\partial H \exp[j(\pi/2 + \theta_z)] \quad (103)$$

$$\bar{C}_n = N \partial\theta_n/\partial H \exp[j(\pi/2 + \theta_n)] \quad (104)$$

$$\bar{C}_a = A \partial\theta_a/\partial H \exp[j(\pi/2 + \theta_a)] \quad (105)$$

Two special cases are of interest.

Case a. Constant Magnetic Field.

This case corresponds to radioactive transmutations or nuclear reactions in a constant magnetic field. For this case equations (97) through (99) give

$$(d\bar{z}/d\bar{t})_H = \cos \beta_{tt} \bar{B}_z dZ/dt \exp[j(\theta_z - \theta_t - \beta_{tt})] \quad (106)$$

$$(d\bar{n}/d\bar{t})_H = \cos \beta_{tt} \bar{B}_n dN/dt \exp[j(\theta_n - \theta_t - \beta_{tt})] \quad (107)$$

$$(d\bar{a}/d\bar{t})_H = \cos \beta_{tt} \bar{B}_a dA/dt \exp[j(\theta_a - \theta_t - \beta_{tt})] \quad (108)$$

For this case the actual size of the nucleus is changing.

Case b. Constant Z, N and A.

This case corresponds to internal phase changes in a fixed nucleus in the presence of a time varying electromagnetic field. Equations (97) through (99) then give

$$\begin{aligned}
 (d\bar{z}/d\bar{t})_Z &= \cos \beta_{tt} \bar{C}_Z dH/dt \exp[j(\theta_Z - \theta_t - \beta_{tt})] \\
 &= \cos \beta_{tt} Z \partial \theta_Z / \partial H dH/dt \exp[j(\pi/2 + 2\theta_Z - \theta_t - \beta_{tt})]
 \end{aligned}
 \tag{109}$$

$$\begin{aligned}
 (d\bar{n}/d\bar{t})_N &= \cos \beta_{tt} \bar{C}_N dH/dt \exp[j(\theta_N - \theta_t - \beta_{tt})] \\
 &= \cos \beta_{tt} N \partial \theta_N / \partial H dH/dt \exp[j(\pi/2 + 2\theta_N - \theta_t - \beta_{tt})]
 \end{aligned}
 \tag{110}$$

$$\begin{aligned}
 (d\bar{a}/d\bar{t})_A &= \cos \beta_{tt} \bar{C}_A dH/dt \exp[j(\theta_A - \theta_t - \beta_{tt})] \\
 &= \cos \beta_{tt} A \partial \theta_A / \partial H dH/dt \exp[j(\pi/2 + 2\theta_A - \theta_t - \beta_{tt})]
 \end{aligned}
 \tag{111}$$

Case b corresponds to the decay of nuclei through internal phase changes associated with a time varying magnetic field, and is considered in detail in Section 3.

The time derivatives of \bar{z} , \bar{n} and \bar{a} given in equations (97) through (99) can also be written in a more general form involving only one exponential term by noting that equations (40) through (42) give

$$\begin{aligned}
 d\bar{z}/d\bar{t} &= \cos \beta_{tt} \sec \beta_{zz} dz/dt \exp(j\phi_{zt}) \\
 &= \cos \beta_{tt} \csc \beta_{zz} z d\theta_Z/dt \exp(j\phi_{zt}) \\
 &= \sin \beta_{tt} \sec \beta_{zz} t^{-1} dz/d\theta_t \exp(j\phi_{zt}) \\
 &= \sin \beta_{tt} \csc \beta_{zz} z/t d\theta_Z/d\theta_t \exp(j\phi_{zt})
 \end{aligned}
 \tag{112}$$

$$\begin{aligned}
 d\bar{n}/d\bar{t} &= \cos \beta_{tt} \sec \beta_{nn} dn/dt \exp(j\phi_{nt}) \\
 &= \cos \beta_{tt} \csc \beta_{nn} n d\theta_N/dt \exp(j\phi_{nt}) \\
 &= \sin \beta_{tt} \sec \beta_{nn} t^{-1} dn/d\theta_t \exp(j\phi_{nt}) \\
 &= \sin \beta_{tt} \csc \beta_{nn} n/t d\theta_N/d\theta_t \exp(j\phi_{nt})
 \end{aligned}
 \tag{113}$$

$$\begin{aligned}
 d\bar{a}/d\bar{t} &= \cos \beta_{tt} \sec \beta_{aa} da/dt \exp(j\phi_{at}) \\
 &= \cos \beta_{tt} \csc \beta_{aa} a d\theta_A/dt \exp(j\phi_{at}) \\
 &= \sin \beta_{tt} \sec \beta_{aa} t^{-1} da/d\theta_t \exp(j\phi_{at}) \\
 &= \sin \beta_{tt} \csc \beta_{aa} a/t d\theta_A/d\theta_t \exp(j\phi_{at})
 \end{aligned}
 \tag{114}$$

where β_{tt} is given by equation (10), and where

$$\tan \beta_{zz} = z \partial \theta_z / \partial z \quad \tan \beta_{nn} = n \partial \theta_n / \partial n \quad \tan \beta_{aa} = a \partial \theta_a / \partial a \quad (115)$$

$$\phi_{zt} = \theta_z + \beta_{zz} - \theta_t - \beta_{tt} \quad (116)$$

$$\phi_{nt} = \theta_n + \beta_{nn} - \theta_t - \beta_{tt} \quad (117)$$

$$\phi_{at} = \theta_a + \beta_{aa} - \theta_t - \beta_{tt} \quad (118)$$

The derivatives dz/dt , dn/dt and da/dt are evaluated using equation (67) as follows

$$dz/dt = dZ/dt \cos \theta_z - Z \sin \theta_z d\theta_z/dt \quad (119)$$

$$= dZ/dt (\cos \theta_z - Z \sin \theta_z \partial \theta_z / \partial Z) - Z \sin \theta_z \partial \theta_z / \partial H dH/dt$$

$$dn/dt = dN/dt \cos \theta_n - N \sin \theta_n d\theta_n/dt \quad (120)$$

$$= dN/dt (\cos \theta_n - N \sin \theta_n \partial \theta_n / \partial N) - N \sin \theta_n \partial \theta_n / \partial H dH/dt$$

$$da/dt = dA/dt \cos \theta_a - A \sin \theta_a d\theta_a/dt \quad (121)$$

$$= dA/dt (\cos \theta_a - A \sin \theta_a \partial \theta_a / \partial A) - A \sin \theta_a \partial \theta_a / \partial H dH/dt$$

These derivatives appear in equations (112) through (114).

Combining equation (67) and equations (112) through (115) gives for the case where both H and Z , N , A are varying

$$d\bar{z}/d\bar{t} = J_z \cos \beta_{tt} \sec \beta_{zz} \cos \theta_z \exp(j\phi_{zt}) \quad (122)$$

$$= I_z \cos \beta_{tt} \csc \beta_{zz} \cos \theta_z \exp(j\phi_{zt})$$

$$= (I_z^2 + J_z^2)^{1/2} \cos \beta_{tt} \cos \theta_z \exp(j\phi_{zt})$$

$$d\bar{n}/d\bar{t} = J_n \cos \beta_{tt} \sec \beta_{nn} \cos \theta_n \exp(j\phi_{nt}) \quad (123)$$

$$= I_n \cos \beta_{tt} \csc \beta_{nn} \cos \theta_n \exp(j\phi_{nt})$$

$$= (I_n^2 + J_n^2)^{1/2} \cos \beta_{tt} \cos \theta_n \exp(j\phi_{nt})$$

$$\begin{aligned}
d\bar{a}/d\bar{t} &= J_a \cos \beta_{tt} \sec \beta_{aa} \cos \theta_a \exp(j\phi_{at}) \\
&= I_a \cos \beta_{tt} \csc \beta_{aa} \cos \theta_a \exp(j\phi_{at}) \\
&= (I_a^2 + J_a^2)^{1/2} \cos \beta_{tt} \cos \theta_a \exp(j\phi_{at})
\end{aligned} \tag{124}$$

where

$$\tan \beta_{zz} = I_z/J_z \quad \tan \beta_{nn} = I_n/J_n \quad \tan \beta_{aa} = I_a/J_a \tag{125}$$

$$I_z = \tan \alpha_{ZZ}^H dZ/dt + Z/H \tan \alpha_{HH}^Z dH/dt \tag{126}$$

$$I_n = \tan \alpha_{NN}^H dN/dt + N/H \tan \alpha_{HH}^N dH/dt \tag{127}$$

$$I_a = \tan \alpha_{AA}^H dA/dt + A/H \tan \alpha_{HH}^A dH/dt \tag{128}$$

$$J_z = (1 - \tan \theta_z \tan \alpha_{ZZ}^H) dZ/dt - Z/H \tan \theta_z \tan \alpha_{HH}^Z dH/dt \tag{129}$$

$$J_n = (1 - \tan \theta_n \tan \alpha_{NN}^H) dN/dt - N/H \tan \theta_n \tan \alpha_{HH}^N dH/dt \tag{130}$$

$$J_a = (1 - \tan \theta_a \tan \alpha_{AA}^H) dA/dt - A/H \tan \theta_a \tan \alpha_{HH}^A dH/dt \tag{131}$$

and where

$$\tan \alpha_{ZZ}^H = Z\partial\theta_z/\partial Z \quad \tan \alpha_{NN}^H = N\partial\theta_n/\partial N \quad \tan \alpha_{AA}^H = A\partial\theta_a/\partial A \tag{132}$$

$$\tan \alpha_{HH}^Z = H\partial\theta_z/\partial H \quad \tan \alpha_{HH}^N = H\partial\theta_n/\partial H \quad \tan \alpha_{HH}^A = H\partial\theta_a/\partial H \tag{133}$$

Two special cases are of interest, for application to nuclear reactions and transmutations and which can be experimentally verified.

Case a. External Radioactive Decay in a Constant Magnetic Field.

For this case it follows from equations (125) through (133) that

$$I_z^H = \tan \alpha_{ZZ}^H dZ/dt \tag{134}$$

$$I_n^H = \tan \alpha_{NN}^H dN/dt \tag{135}$$

$$I_a^H = \tan \alpha_{AA}^H dA/dt \tag{136}$$

$$J_z^H = (1 - \tan \theta_z \tan \alpha_{zz}^H) dZ/dt \quad (137)$$

$$J_n^H = (1 - \tan \theta_n \tan \alpha_{nn}^H) dN/dt \quad (138)$$

$$J_a^H = (1 - \tan \theta_a \tan \alpha_{aa}^H) dA/dt \quad (139)$$

$$\tan \beta_{zz}^H = I_z^H/J_z^H = \tan \alpha_{zz}^H (1 - \tan \theta_z \tan \alpha_{zz}^H)^{-1} \quad (140)$$

$$\tan \beta_{nn}^H = I_n^H/J_n^H = \tan \alpha_{nn}^H (1 - \tan \theta_n \tan \alpha_{nn}^H)^{-1} \quad (141)$$

$$\tan \beta_{aa}^H = I_a^H/J_a^H = \tan \alpha_{aa}^H (1 - \tan \theta_a \tan \alpha_{aa}^H)^{-1} \quad (142)$$

$$\alpha_{HH}^Z = \pi/2 \quad \alpha_{HH}^N = \pi/2 \quad \alpha_{HH}^A = \pi/2 \quad (143)$$

From equations (106) through (108) and (122) through (143) it follows for constant H that

$$\begin{aligned} (d\bar{z}/d\bar{t})_H &= J_z^H \cos \beta_{tt} \sec \beta_{zz}^H \cos \theta_z \exp(j\phi_{zt}^H) \\ &= I_z^H \cos \beta_{tt} \csc \beta_{zz}^H \cos \theta_z \exp(j\phi_{zt}^H) \\ &= [(I_z^H)^2 + (J_z^H)^2]^{1/2} \cos \beta_{tt} \cos \theta_z \exp(j\phi_{zt}^H) \end{aligned} \quad (144)$$

$$\begin{aligned} (d\bar{n}/d\bar{t})_H &= J_n^H \cos \beta_{tt} \sec \beta_{nn}^H \cos \theta_n \exp(j\phi_{nt}^H) \\ &= I_n^H \cos \beta_{tt} \csc \beta_{nn}^H \cos \theta_n \exp(j\phi_{nt}^H) \\ &= [(I_n^H)^2 + (J_n^H)^2]^{1/2} \cos \beta_{tt} \cos \theta_n \exp(j\phi_{nt}^H) \end{aligned} \quad (145)$$

$$\begin{aligned} (d\bar{a}/d\bar{t})_H &= J_a^H \cos \beta_{tt} \sec \beta_{aa}^H \cos \theta_a \exp(j\phi_{at}^H) \\ &= I_a^H \cos \beta_{tt} \csc \beta_{aa}^H \cos \theta_a \exp(j\phi_{at}^H) \\ &= [(I_a^H)^2 + (J_a^H)^2]^{1/2} \cos \beta_{tt} \cos \theta_a \exp(j\phi_{at}^H) \end{aligned} \quad (146)$$

where

$$\phi_{zt}^H = \theta_z + \beta_{zz}^H - \theta_t - \beta_{tt} \quad (147)$$

$$\phi_{nt}^H = \theta_n + \beta_{nn}^H - \theta_t - \beta_{tt} \quad (148)$$

$$\phi_{at}^H = \theta_a + \beta_{aa}^H - \theta_t - \beta_{tt} \quad (149)$$

Simple algebra shows that equations (106) through (108) are equivalent to equations (144) through (146).

Case b. Internal Phase Radioactive Decay due to a Time Dependent Magnetic Field.

This case corresponds to constant values of Z, N and A. Equations (125) through (133) give

$$\tan \beta_{zz}^Z = -\cot \theta_z \quad \beta_{zz}^Z = \theta_z + \pi/2 \quad (150)$$

$$\tan \beta_{nn}^N = -\cot \theta_n \quad \beta_{nn}^N = \theta_n + \pi/2 \quad (151)$$

$$\tan \beta_{aa}^A = -\cot \theta_a \quad \beta_{aa}^A = \theta_a + \pi/2 \quad (152)$$

$$\csc \beta_{zz}^Z = \sec \theta_z \quad \csc \beta_{nn}^N = \sec \theta_n \quad \csc \beta_{aa}^A = \sec \theta_a \quad (153)$$

$$\sec \beta_{zz}^Z = -\csc \theta_z \quad \sec \beta_{nn}^N = -\csc \theta_n \quad \sec \beta_{aa}^A = -\csc \theta_a \quad (154)$$

$$\alpha_{ZZ}^H = \pi/2 \quad \alpha_{NN}^H = \pi/2 \quad \alpha_{AA}^H = \pi/2 \quad (155)$$

$$I_z^Z = Z/H \tan \alpha_{HH}^Z dH/dt \quad (156)$$

$$I_n^N = N/H \tan \alpha_{HH}^N dH/dt \quad (157)$$

$$I_a^A = A/H \tan \alpha_{HH}^A dH/dt \quad (158)$$

$$J_z^Z = -Z/H \tan \theta_z \tan \alpha_{HH}^Z dH/dt \quad (159)$$

$$J_n^N = -N/H \tan \theta_n \tan \alpha_{HH}^N dH/dt \quad (160)$$

$$J_a^A = -A/H \tan \theta_a \tan \alpha_{HH}^A dH/dt \quad (161)$$

For constant Z, N and A it follows that

$$[(I_z^Z)^2 + (J_z^Z)^2]^{1/2} = Z/H \tan \alpha_{HH}^Z \sec \theta_z dH/dt \quad (162)$$

$$= Z \partial \theta_z / \partial H \sec \theta_z dH/dt$$

$$[(I_n^N)^2 + (J_n^N)^2]^{1/2} = N/H \tan \alpha_{HH}^N \sec \theta_n dH/dt \quad (163)$$

$$= N \partial \theta_n / \partial H \sec \theta_n dH/dt$$

$$[(I_a^A)^2 + (J_a^A)^2]^{1/2} = A/H \tan \alpha_{HH}^A \sec \theta_a dH/dt \quad (164)$$

$$= A \partial \theta_a / \partial H \sec \theta_a dH/dt$$

Combining equations (122) through (124) and equations (162) through (164) gives for Z, N and A fixed

$$(d\bar{z}/d\bar{t})_Z = \cos \beta_{tt} Z(d\theta_z/dt)_Z \exp[j(\pi/2 + 2\theta_z - \theta_t - \beta_{tt})] \quad (165)$$

$$= \cos \beta_{tt} Z \partial \theta_z / \partial H dH/dt \exp[j(\pi/2 + 2\theta_z - \theta_t - \beta_{tt})]$$

$$(d\bar{n}/d\bar{t})_N = \cos \beta_{tt} N(d\theta_n/dt)_N \exp[j(\pi/2 + 2\theta_n - \theta_t - \beta_{tt})] \quad (166)$$

$$= \cos \beta_{tt} N \partial \theta_n / \partial H dH/dt \exp[j(\pi/2 + 2\theta_n - \theta_t - \beta_{tt})]$$

$$(d\bar{a}/d\bar{t})_A = \cos \beta_{tt} A(d\theta_a/dt)_A \exp[j(\pi/2 + 2\theta_a - \theta_t - \beta_{tt})] \quad (167)$$

$$= \cos \beta_{tt} A \partial \theta_a / \partial H dH/dt \exp[j(\pi/2 + 2\theta_a - \theta_t - \beta_{tt})]$$

Internal phase radioactive decay in a time varying electromagnetic field is treated in Section 3.

C. Dependence of the Internal Phase Angles on Z, N, A and H.

The following material is a simple way of relating the internal space phase angles θ_z , θ_n and θ_a to the strength of an applied electromagnetic field. It has been assumed that the complex atomic number \bar{z} , neutron number \bar{n} and atomic mass number \bar{a} arise from solutions of the azimuthal equations (44) through (46) so that θ_z , θ_n and θ_a are given by equation (68). Then if it is assumed that the internal space phase angles $\theta_{\phi z}$, $\theta_{\phi n}$ and $\theta_{\phi a}$ of the complex number azimuthal angles $\bar{\phi}_z$, $\bar{\phi}_n$ and $\bar{\phi}_a$ corresponding to \bar{z} , \bar{n} and \bar{a} as given in equation (43) are proportional to the internal phase angles θ_α of the complex number cartesian coordinates $\bar{\alpha}$, where $\alpha = x, y, z$ with $\theta_\alpha = \theta_x = \theta_y = \theta_z$, it follows that

$$\theta_z = -\theta_{\phi z} = -\omega_z \theta_\alpha \quad (168)$$

$$\theta_n = -\theta_{\phi n} = -\omega_n \theta_\alpha \quad (169)$$

$$\theta_a = -\theta_{\phi a} = -\omega_a \theta_\alpha \quad (170)$$

where ω_z , ω_n and ω_a = constants whose values may be taken to be of the order of unity as a first approximation. The internal space phase angles of the complex number cartesian coordinates are related to the internal space phase angle of a magnetic field by⁵⁷

$$\theta_H = -\theta_\alpha - \theta_t \quad (171)$$

For free particles the relation $\theta_\alpha = 2\theta_t$ is valid so that

$$\theta_H = -3\theta_t = -3/2\theta_\alpha \quad (172)$$

and therefore

$$\theta_z = -\theta_{\phi z} = 2/3\omega_z \theta_H \equiv \alpha_z(Z,A)\theta_H \quad (173)$$

$$\theta_n = -\theta_{\phi n} = 2/3\omega_n \theta_H \equiv \alpha_n(Z,A)\theta_H \quad (174)$$

$$\theta_a = -\theta_{\phi a} = 2/3\omega_a \theta_H \equiv \alpha_a(Z,A)\theta_H \quad (175)$$

At the fission condition, equations (473), (487) and (488) give

$$\alpha_a(Z,A) \sim 2\alpha_z(Z,A) \quad (176)$$

The constants $\alpha_z(Z,A)$, $\alpha_n(Z,A)$ and $\alpha_a(Z,A)$ are of the order of unity to a first approximation so that at incipient fission

$$\theta_z \sim \theta_H \quad \theta_n \sim \theta_H \quad \theta_a \sim 2\theta_H \quad (177)$$

and in general to a first approximation near the valley of beta stability

$$\theta_z \sim \theta_H \quad \theta_n \sim \theta_H \quad \theta_a \sim \theta_H \quad (178)$$

Therefore the internal phase angles of the atomic number, neutron number and atomic mass number can be controlled by applying an electromagnetic field whose internal phase angle of the magnetic field component θ_H is determined by the strength of the applied electromagnetic field.

It is expected that the value of θ_H will depend on the field strength of the applied electromagnetic field and can be represented as a power series as follows

$$\theta_H(H) = \theta_H(0) + h_1 H + h_2 H^2 + \dots \quad (179)$$

For practical calculations it will be assumed that $\theta_H(0) \sim 0$. Combining equations (173) through (175) with equation (179) gives

$$\theta_z = \theta_{zo} + \alpha_z (h_1 H + h_2 H^2 + \dots) \quad (180)$$

$$\theta_n = \theta_{no} + \alpha_n (h_1 H + h_2 H^2 + \dots) \quad (181)$$

$$\theta_a = \theta_{ao} + \alpha_a (h_1 H + h_2 H^2 + \dots) \quad (182)$$

where

$$\theta_{zo} = \alpha_z \theta_H(0) \quad \theta_{no} = \alpha_n \theta_H(0) \quad \theta_{ao} = \alpha_a \theta_H(0) \quad (183)$$

where θ_{zo} , θ_{no} and θ_{ao} = intrinsic phase angles of the atomic number, neutron number and atomic mass number for zero value of the applied electromagnetic field. For practical calculations

$$\theta_{zo} \sim 0 \quad \theta_{no} \sim 0 \quad \theta_{ao} \sim 0 \quad (184)$$

The magnetic field strength H can refer to either a static magnetic field or the magnetic field component of an electromagnetic wave such as in the case of γ rays.

The following nuclear magnetic internal phase angle compliance coefficients can be defined

$$\begin{aligned} C_{\theta z}^H(Z, A) &= \partial \theta_z / \partial H = \alpha_z (h_1 + 2h_2 H + 3h_3 H^2 + \dots) \\ &= C_{\theta zo}^H + C_{\theta z1}^H H + C_{\theta z2}^H H^2 + \dots \end{aligned} \quad (185)$$

$$\begin{aligned} C_{\theta n}^H(Z, A) &= \partial \theta_n / \partial H = \alpha_n (h_1 + 2h_2 H + 3h_3 H^2 + \dots) \\ &= C_{\theta no}^H + C_{\theta n1}^H H + C_{\theta n2}^H H^2 + \dots \end{aligned} \quad (186)$$

$$\begin{aligned} C_{\theta a}^H(Z, A) &= \partial \theta_a / \partial H = \alpha_a (h_1 + 2h_2 H + 3h_3 H^2 + \dots) \\ &= C_{\theta ao}^H + C_{\theta a1}^H H + C_{\theta a2}^H H^2 + \dots \end{aligned} \quad (187)$$

where

$$C_{\theta zo}^H = \alpha_z h_1 \quad C_{\theta z1}^H = 2\alpha_z h_2 \quad C_{\theta z2}^H = 3\alpha_z h_3 \quad (188)$$

$$C_{\theta no}^H = \alpha_n h_1 \quad C_{\theta n1}^H = 2\alpha_n h_2 \quad C_{\theta n2}^H = 3\alpha_n h_3 \quad (189)$$

$$C_{\theta ao}^H = \alpha_a h_1 \quad C_{\theta a1}^H = 2\alpha_a h_2 \quad C_{\theta a2}^H = 3\alpha_a h_3 \quad (190)$$

Combining equations (180) through (182) with equations (188) through (190) gives

$$\theta_z = \theta_{zo} + C_{\theta zo}^H H + 1/2 C_{\theta z1}^H H^2 + 1/3 C_{\theta z2}^H H^3 + \dots \quad (191)$$

$$\theta_n = \theta_{no} + C_{\theta no}^H H + 1/2 C_{\theta n1}^H H^2 + 1/3 C_{\theta n2}^H H^3 + \dots \quad (192)$$

$$\theta_a = \theta_{ao} + C_{\theta ao}^H H + 1/2 C_{\theta a1}^H H^2 + 1/3 C_{\theta a2}^H H^3 + \dots \quad (193)$$

where $\theta_{zo} \sim 0$, $\theta_{no} \sim 0$ and $\theta_{ao} \sim 0$ and can generally be neglected.

Equations (191) through (193) can be combined with equations (97) through (111) to calculate $d\bar{z}/d\bar{t}$, $d\bar{n}/d\bar{t}$ and $d\bar{a}/d\bar{t}$ in terms of the applied electromagnetic field. Combining equations (132) and (133) with equations (191) through (193) gives

$$\tan \alpha_{ZZ}^H = Z \partial \theta_{zo} / \partial Z + Z \partial C_{\theta zo}^H / \partial Z H + 1/2 Z \partial C_{\theta z1}^H / \partial Z H^2 + \dots \quad (194)$$

$$\tan \alpha_{NN}^H = N \partial \theta_{no} / \partial N + N \partial C_{\theta no}^H / \partial N H + 1/2 N \partial C_{\theta n1}^H / \partial N H^2 + \dots \quad (195)$$

$$\tan \alpha_{AA}^H = A \partial \theta_{ao} / \partial A + A \partial C_{\theta ao}^H / \partial A H + 1/2 A \partial C_{\theta a1}^H / \partial A H^2 + \dots \quad (196)$$

$$\tan \alpha_{HH}^Z = H C_{\theta z}^H = H (C_{\theta zo}^H + C_{\theta z1}^H H + C_{\theta z2}^H H^2 + \dots) \quad (197)$$

$$\tan \alpha_{HH}^N = H C_{\theta n}^H = H (C_{\theta no}^H + C_{\theta n1}^H H + C_{\theta n2}^H H^2 + \dots) \quad (198)$$

$$\tan \alpha_{HH}^A = H C_{\theta a}^H = H (C_{\theta ao}^H + C_{\theta a1}^H H + C_{\theta a2}^H H^2 + \dots) \quad (199)$$

The expressions in equations (194) through (199) can be inserted into equations (122) through (167) to calculate $d\bar{z}/d\bar{t}$, $d\bar{n}/d\bar{t}$ and $d\bar{a}/d\bar{t}$ in terms of a strength parameter H of the applied electromagnetic field. For example, in the case of nuclei with fixed values of Z , N and A the effect of a changing magnetic field is obtained from equations (150) through (152) and (165) through (167) to be

$$(\partial \bar{z} / \partial H)_Z = Z (\partial \theta_z / \partial H)_Z \exp[j(2\theta_z + \pi/2)] = Z C_{\theta z}^H \exp[j(\theta_z + \beta_{zz}^Z)] \quad (200)$$

$$(\partial \bar{n} / \partial H)_N = N (\partial \theta_n / \partial H)_N \exp[j(2\theta_n + \pi/2)] = N C_{\theta n}^H \exp[j(\theta_n + \beta_{nn}^N)] \quad (201)$$

$$(\partial \bar{a} / \partial H)_A = A (\partial \theta_a / \partial H)_A \exp[j(2\theta_a + \pi/2)] = A C_{\theta a}^H \exp[j(\theta_a + \beta_{aa}^A)] \quad (202)$$

The previous equations are all exact.

For simple estimations of the effects of an electromagnetic field, the following simple linear approximations can be made by dropping all nonlinear effects of the electromagnetic field in the magnetic compliance coefficients in equations (191) through (193)

$$\theta_z \sim C_{\theta zo}^H H \quad \theta_n \sim C_{\theta no}^H H \quad \theta_a \sim C_{\theta ao}^H H \quad (203)$$

It is often convenient to invert equation (203) as

$$H \sim K_{\theta zo}^H \theta_z \quad H \sim K_{\theta no}^H \theta_n \quad H \sim K_{\theta ao}^H \theta_a \quad (204)$$

where the nuclear magnetic internal angle stiffness coefficients are given by

$$K_{\theta zo}^H = (C_{\theta zo}^H)^{-1} \quad K_{\theta no}^H = (C_{\theta no}^H)^{-1} \quad K_{\theta ao}^H = (C_{\theta ao}^H)^{-1} \quad (205)$$

Generally it is more convenient to use the magnetic induction field B in calculations so that the following corresponding nuclear magnetic compliance and stiffness coefficients are defined by

$$\theta_z \sim C_{\theta zo}^B B \quad \theta_n \sim C_{\theta no}^B B \quad \theta_a \sim C_{\theta ao}^B B \quad (206)$$

and their inverses

$$B \sim K_{\theta zo}^B \theta_z \quad B \sim K_{\theta no}^B \theta_n \quad B \sim K_{\theta ao}^B \theta_a \quad (207)$$

where

$$K_{\theta zo}^B = (C_{\theta zo}^B)^{-1} \quad K_{\theta no}^B = (C_{\theta no}^B)^{-1} \quad K_{\theta ao}^B = (C_{\theta ao}^B)^{-1} \quad (208)$$

In most cases the higher order terms in equations (185) through (187) and (191) through (193) can be ignored so that the subscript "o" can be dropped and the following are taken to be exact relationships

$$\theta_z = C_{\theta z}^H H \quad \theta_n = C_{\theta n}^H H \quad \theta_a = C_{\theta a}^H H \quad (209)$$

$$H = K_{\theta z}^H \theta_z \quad H = K_{\theta n}^H \theta_n \quad H = K_{\theta a}^H \theta_a \quad (210)$$

$$\theta_z = C_{\theta z}^B B \quad \theta_n = C_{\theta n}^B B \quad \theta_a = C_{\theta a}^B B \quad (211)$$

$$B = K_{\theta z}^B \theta_z \quad B = K_{\theta n}^B \theta_n \quad B = K_{\theta a}^B \theta_a \quad (212)$$

where

$$K_{\theta z}^H = (C_{\theta z}^H)^{-1} \quad K_{\theta n}^H = (C_{\theta n}^H)^{-1} \quad K_{\theta a}^H = (C_{\theta a}^H)^{-1} \quad (213)$$

$$K_{\theta z}^B = (C_{\theta z}^B)^{-1} \quad K_{\theta n}^B = (C_{\theta n}^B)^{-1} \quad K_{\theta a}^B = (C_{\theta a}^B)^{-1} \quad (214)$$

3. RADIOACTIVE DECAY OF ATOMIC NUCLEI IN AN ELECTROMAGNETIC FIELD. This section considers the radioactive decay of an assemblage of identical radioactive nuclei located in an electromagnetic field, or other external field such as gravity, which induces internal phase angles to the number of nuclei and the total number of constituent nucleons contained within the nuclei and to the atomic mass number of each nucleus. The addition law is developed for the complex total nucleon number, complex number of atomic nuclei and complex atomic mass number.

A. Addition Law for Complex Total Nucleon Number, Complex Number of Atomic Nuclei and Complex Atomic Mass Number.

Consider a system of atomic nuclei in an external electromagnetic field. The presence of the electromagnetic field requires that particle number be represented as a complex number in an internal space. Therefore the complex number of total nucleons (protons and neutrons within the nuclei), the complex number of atomic nuclei and the complex atomic mass number are written as

$$\bar{N}_n = N_n \exp(j\theta_{Nn}) \quad (215)$$

$$\bar{N} = N \exp(j\theta_N) \quad (216)$$

$$\bar{a} = a \exp(j\theta_a) \quad (217)$$

where \bar{N}_n , N_n and θ_{Nn} = complex number value, magnitude and internal phase angle of the total number of nucleons situated within all of the atomic nuclei; \bar{N} , N and θ_N = complex number value, magnitude and internal phase angle of the number of atomic nuclei; and as before \bar{a} , a and θ_a = complex number value, magnitude and internal phase angle of the atomic mass number of each nucleus. In analogy to equation (67) the magnitudes that appear in equations (215) through (217) are written as

$$N_n = \eta_n \cos \theta_{Nn} \quad (218)$$

$$N = \eta \cos \theta_N \quad (219)$$

$$a = A \cos \theta_a \quad (220)$$

where η_n = integer number of total number of nucleons within the atomic nuclei, η = integer number of atomic nuclei, and as before A = atomic mass number which by definition is an integer. These integer numbers satisfy the equation

$$\eta_n = \eta A \quad (221)$$

which represents the fundamental law of baryon number conservation which is universally valid. The measured nucleon number, nuclei number and atomic mass number are given by the real parts of equations (215) through (217) respectively

$$N_{nm} = \eta_n \cos^2 \theta_{Nn} \quad N_m = \eta \cos^2 \theta_N \quad a_m = A \cos^2 \theta_a \quad (222)$$

and are not integers.

In analogy to equation (82) which applies to a single nucleus the addition law for the complex total nucleon number is given by

$$W' + \bar{N}_n = \eta \bar{a} \quad (223)$$

which is subject to the validity of equation (221). The component equations of equation (223) are

$$W' + \eta_n \cos^2 \theta_{Nn} = \eta A \cos^2 \theta_a \quad (224)$$

$$\eta_n \cos \theta_{Nn} \sin \theta_{Nn} = \eta A \cos \theta_a \sin \theta_a \quad (225)$$

which are two equations for the two unknown quantities W' and θ_{Nn} . It is easy to show that the solutions to equations (224) and (225) are

$$\cos^2 \theta_{Nn} = 1/2[1 + (1 - 4f'^2)^{1/2}] \quad (226)$$

$$\begin{aligned} W' &= \eta A (\cos^2 \theta_a - \cos^2 \theta_{Nn}) \\ &= \eta A \{ \cos^2 \theta_a - 1/2[1 + (1 - f'^2)^{1/2}] \} \end{aligned} \quad (227)$$

where

$$f' = \cos \theta_a \sin \theta_a \quad (228)$$

where θ_a is assumed to be a known function of the applied electromagnetic field strength. The internal phase angle θ_{Nn} of the number of atomic nuclei is also assumed to be a known function of the external electromagnetic field strength. Equations (218) through (220) can be used to determine N_n , N and a in terms of the electromagnetic field strength H . The baryon number conservation equation (221) is a universal law which is valid whether or not an electromagnetic field is present.

B. Radioactive Decay of Nuclei in the Presence of an Electromagnetic Field.

Radioactive decay of heavy elements has been studied for many years.^{58,59} This section considers the radioactive decay of heavy elements located in an external electromagnetic field. The generalization of the standard radioactive decay law for elements is written as

$$d\bar{N}/d\bar{t} = -\bar{\lambda}\bar{N} \quad (229)$$

where \bar{N} = complex number of atomic nuclei and $\bar{\lambda}$ = complex number radioactive decay constant which can be written as

$$\bar{\lambda} = \lambda \exp(j\theta_\lambda) \quad (230)$$

From equations (216) and (219) it follows that \bar{N} can be written as

$$\bar{N} = \eta \cos \theta_{Nn} \exp(j\theta_{Nn}) \quad N = \eta \cos \theta_{Nn} \quad (231)$$

The time derivative of equation (231) can be written as

$$d\bar{N}/d\bar{t} = \cos \beta_{tt} \sec \beta_{NN} dN/dt \exp(j\phi_{Nt}) \quad (232)$$

$$= \cos \beta_{tt} \csc \beta_{NN} N d\theta_N/dt \exp(j\phi_{Nt}) \quad (233)$$

$$= \sin \beta_{tt} \sec \beta_{NN} t^{-1} dN/d\theta_t \exp(j\phi_{Nt}) \quad (234)$$

$$= \sin \beta_{tt} \csc \beta_{NN} N/t d\theta_N/d\theta_t \exp(j\phi_{Nt}) \quad (235)$$

where

$$\tan \beta_{NN} = N \partial \theta_N / \partial N \quad (236)$$

$$\phi_{Nt} = \theta_N + \beta_{NN} - \theta_t - \beta_{tt} \quad (237)$$

Then the law of radioactive decay of elements given in equation (229) can be written in any of the following forms

$$\cos \beta_{tt} \sec \beta_{NN} dN/dt = -\lambda N \quad (238)$$

$$\cos \beta_{tt} \csc \beta_{NN} d\theta_N/dt = -\lambda \quad (239)$$

$$\sin \beta_{tt} \sec \beta_{NN} t^{-1} dN/d\theta_t = -\lambda N \quad (240)$$

$$\sin \beta_{tt} \csc \beta_{NN} t^{-1} d\theta_N/d\theta_t = -\lambda \quad (241)$$

combined with the following phase angle relationship

$$\phi_{Nt} = \theta_\lambda + \theta_N \quad (242)$$

Combining equations (237) and (242) gives

$$\beta_{NN} - \theta_t - \beta_{tt} = \theta_\lambda \quad (243)$$

The derivative $d\theta_N/dt$ that appears in equation (233) is written as

$$d\theta_N/dt = \partial \theta_N / \partial \eta d\eta/dt + \partial \theta_N / \partial H dH/dt \quad (244)$$

whereas the derivative dN/dt that appears in equation (232) is obtained from equation (231) to be

$$\begin{aligned} dN/dt &= d\eta/dt \cos \theta_N - \eta \sin \theta_N d\theta_N/dt \\ &= d\eta/dt (\cos \theta_N - \eta \sin \theta_N \partial \theta_N / \partial \eta) - \eta \sin \theta_N \partial \theta_N / \partial H dH/dt \end{aligned} \quad (245)$$

Equations (238) through (243) give the general forms of the law of radioactive decay of elements in the presence of an external field.

An additional form of the law of radioactive decay can be obtained by noting that in an analogous fashion to equations (122) through (124) the time

derivative $d\bar{N}/d\bar{t}$ that is required in equation (229) can be obtained from equations (232), (233), (244) and (245) to be

$$d\bar{N}/d\bar{t} = \cos \beta_{tt} \sec \beta_{NN} \cos \theta_N J_N \exp(j\phi_{Nt}) \quad (246)$$

$$= \cos \beta_{tt} \csc \beta_{NN} \cos \theta_N I_N \exp(j\phi_{Nt}) \quad (247)$$

$$= \cos \beta_{tt} \cos \theta_N (I_N^2 + J_N^2)^{1/2} \exp(j\phi_{Nt}) \quad (248)$$

where

$$I_N = \tan \alpha_{\eta\eta}^H d\eta/dt + \eta/H \tan \alpha_{HH}^\eta dH/dt \quad (249)$$

$$J_N = (1 - \tan \theta_N \tan \alpha_{\eta\eta}^H) d\eta/dt - \eta/H \tan \theta_N \tan \alpha_{HH}^\eta dH/dt \quad (250)$$

$$\tan \beta_{NN} = N \partial \theta_N / \partial N = I_N / J_N \quad (251)$$

$$\tan \alpha_{\eta\eta}^H = \eta \partial \theta_N / \partial \eta \quad (252)$$

$$\tan \alpha_{HH}^\eta = H \partial \theta_N / \partial H \quad (253)$$

Then equation (229) can be written as

$$\begin{aligned} & \cos \beta_{tt} \cos \theta_N (I_N^2 + J_N^2)^{1/2} \exp(j\phi_{Nt}) \\ & = -\lambda \eta \cos \theta_N \exp[j(\theta_\lambda + \theta_N)] \end{aligned} \quad (254)$$

which gives the radioactive decay law as

$$\cos \beta_{tt} (I_N^2 + J_N^2)^{1/2} = -\lambda \eta \quad (255)$$

and equations (242) or (243) for the phase angle relationship. Equation (255) is equivalent to equation (238) as can be seen from equations (245) through (248).

At this point the standard law of radioactive decay can be regained when $\theta_N = \text{constant}$, $\theta_t = 0$ and $\beta_{tt} = 0$ in which case equations (229), (232), (245) and (254) reduce to

$$\cos \theta_N d\eta/dt \exp(j\theta_N) = -\lambda \eta \cos \theta_N \exp[j(\theta_N + \theta_\lambda)] \quad (256)$$

which gives the standard law of radioactive decay^{58, 59}

$$d\eta/dt = -\lambda \eta \quad \theta_\lambda = 0 \quad (257)$$

which corresponds to incoherent radioactive decay of atomic nuclei. This limiting case can also be obtained from equations (246) through (255) by noting that

$$\beta_{NN} = 0 \quad \alpha_{nn}^H = 0 \quad \alpha_{HH}^\eta = 0 \quad (258)$$

$$J_N = d\eta/dt \quad I_N = 0 \quad (259)$$

and equations (255) and (243) become equation (257) which is the standard equation for incoherent radioactive decay.

For the case of a static magnetic field, equations (249), (250) and (255) reduce respectively to

$$I_N^H = \tan \alpha_{nn}^H d\eta/dt \quad (260)$$

$$J_N^H = (1 - \tan \theta_N \tan \alpha_{nn}^H) d\eta/dt \quad (261)$$

$$Y \cos \beta_{tt} d\eta/dt = -\lambda \eta \quad (262)$$

where

$$Y = [\tan^2 \alpha_{nn}^H + (1 - \tan \theta_N \tan \alpha_{nn}^H)^2]^{1/2} \quad (263)$$

Equation (262) can be rewritten as

$$d\eta/dt = -\lambda' \eta$$

where the effective decay constant is given by

$$\lambda' = \lambda Y^{-1} \sec \beta_{tt} \quad (265)$$

Therefore the effective radioactive decay constant is predicted to be an increasing function of the strength of an applied static magnetic field. The phase angle condition for this special case is obtained by first noting that from equation (251) it follows that for $H = \text{constant}$

$$\tan \beta_{NN}^H = I_N^H/J_N^H = \tan \alpha_{nn}^H (1 - \tan \theta_N \tan \alpha_{nn}^H)^{-1} \quad (266)$$

and equation (243) gives the phase angle condition as

$$\theta_\lambda = \beta_{NN}^H - \theta_t - \beta_{tt} \quad (267)$$

where θ_λ = internal phase angle of the radioactive decay constant for atomic nuclei in a constant magnetic field.

C. Coherent Radioactive Decay of Atomic Nuclei.

Consider now the case when $\eta = \text{constant}$ which corresponds to a radioactive decay where the integer number of atomic nuclei remains constant and only the internal phase angle θ_N changes due to the presence of a time dependent electromagnetic field. For $\eta = \text{constant}$ equations (249) and (250) give

$$I_N^\eta = \eta/H \tan \alpha_{HH}^\eta dH/dt \quad (268)$$

$$J_N^\eta = -\eta/H \tan \theta_N \tan \alpha_{HH}^\eta dH/dt \quad (269)$$

while equation (251) gives for $\eta = \text{constant}$

$$\tan \beta_{NN}^\eta = -\cot \theta_N \quad \beta_{NN}^\eta = \pi/2 + \theta_N \quad (270)$$

Then equations (253), (255), (268) and (269) and equations (243) and (270) become the following internal phase radioactive decay equations

$$\cos \beta_{tt} \sec \theta_N d\theta_N/dt = -\lambda \quad (271)$$

$$\begin{aligned} \theta_\lambda &= \pi/2 + \theta_N - \theta_t - \beta_{tt} \\ &= \beta_{NN}^\eta - \theta_t - \beta_{tt} \end{aligned} \quad (272)$$

where β_{tt} is given by equation (10). Equations (271) and (272) can also be derived directly from equations (229) and (231) by noting that for $\eta = \text{constant}$

$$\begin{aligned} d\bar{N}/d\bar{t} &= \eta \cos \beta_{tt} d\theta_N/dt \exp[j(\pi/2 + 2\theta_N - \theta_t - \beta_{tt})] \\ &= \bar{N} \cos \beta_{tt} \sec \theta_N d\theta_N/dt \exp[j(\pi/2 + \theta_N - \theta_t - \beta_{tt})] \end{aligned} \quad (273)$$

and

$$\begin{aligned} d\bar{N} &= n d\theta_N \exp[j(\pi/2 + 2\theta_N)] & |d\bar{N}| &= n d\theta_N \\ &= j n d\theta_N \exp(j2\theta_N) \\ &= j \bar{N} \sec \theta_N d\theta_N \exp(j\theta_N) \\ &= \bar{N} \sec \theta_N d\theta_N \exp[j(\pi/2 + \theta_N)] \end{aligned} \quad (274)$$

Equation (274) represents the coherent form of change of the special type of complex number given in equations (216) and (219). Equations (271) and (272) are coupled simultaneous differential equations that determine θ_N and θ_t . Thus for slowly changing values of θ_t equation (272) can be written as

$$\theta_\lambda = \pi/2 + \theta_N - \theta_t - t \partial \theta_t / \partial t \quad (275)$$

which is an approximate equation.

The solution to the coupled internal phase radioactive decay equations (271) and (272) is in general not simply obtained and can only be done exactly by numerical methods on a computer. Some insight can be obtained, however, by

assuming that $\theta_t(t)$ is a known function and solving equation (271) for θ_N . Simple integration of equation (271) gives

$$(\sec \theta_N + \tan \theta_N)/(\sec \theta_N^0 + \tan \theta_N^0) = \exp[-\lambda g(t)] \quad (276)$$

where

$$g(t) = \int_0^t \sec \beta_{tt} dt \quad (277)$$

and where $\theta_N = \theta_N^0$ for $t = 0$ is an initial condition. For $t \rightarrow \infty$ it will be assumed that $g(t) \rightarrow \infty$ so that for this limit

$$\sec \theta_N^\infty + \tan \theta_N^\infty = 0 \quad (278)$$

which gives

$$\theta_N^\infty = -\pi/2 \quad (279)$$

From equations (272) and (275) it follows that the asymptotic values of θ_t are given for $t \rightarrow \infty$ by

$$\begin{aligned} \theta_\lambda &= -\theta_t - \beta_{tt} \\ \sim -\theta_t - t \partial \theta_t / \partial t \end{aligned} \quad (280)$$

So that for $t \rightarrow \infty$ the following approximate equation is valid

$$\theta_t = -\theta_\lambda + c/t \quad (281)$$

where $c = \text{constant}$. Therefore for $t = \infty$ it follows that $\theta_t = -\theta_\lambda$. Therefore for a system of radioactive nuclei that is decaying by changes in the internal phase angle of the nuclei number for a fixed integer number of nuclei, the equilibrium value of θ_t that is obtained after a long period of time is given by $\theta_t = -\theta_\lambda$. Chemical reactions obeying the mass action law can also occur by changes in the internal phase angles of the reactant species numbers.

4. NUCLEAR MASS FORMULA FOR ATOMIC NUCLEI IN AN EXTERNAL ELECTROMAGNETIC OR GRAVITATIONAL FIELD. Nuclear mass formulas are important tools for investigating nuclear structure and nuclear processes such as fission and the neutron and proton capture by atomic nuclei.⁶⁰⁻⁶⁷ The Weizsäcker-Bethe liquid drop mass formula was developed years ago to represent nuclear masses by combined classical and quantum mechanical techniques.⁶⁰⁻⁶⁷ This section develops a broken symmetry form of the liquid drop nuclear mass formula that is required to investigate the clean fission of lighter than actinide nuclei by thermal neutrons in the presence of a γ ray field that is considered in Sections 5 through 7.

A. Complex Number Nuclear Radius for Subactinide Nuclei.

The concept of a nuclear radius enters directly into the derivation of

the liquid drop type of nuclear mass formula. The simplest nuclear radius formula is⁶⁰⁻⁶⁷

$$R = bA^{1/3} \quad b = 1.2 \times 10^{-15} \text{ m} = 1.2 \text{ fm} \quad (282)$$

In an external field the nuclear radius must be represented by a complex number in an internal space. The complex number generalization of equation (282) is written as

$$\bar{R} = R \exp(j\theta_R) = \bar{b}\bar{a}^{1/3} \quad (283)$$

where $\bar{b} = b \exp(j\theta_b)$ = complex number constant and \bar{a} = complex number atomic mass number given by equation (42). Combining equations (217), (220) and (283) gives for subactinide nuclei

$$\bar{R} = bA^{1/3} \cos^{1/3}\theta_a \exp[j(\theta_b + \theta_a/3)] \quad (284)$$

$$R = bA^{1/3} \cos^{1/3}\theta_a \quad \theta_R = \theta_b + \theta_a/3 \quad (285)$$

The measured nuclear radius is given by the real part of equation (284)

$$R_m = R \cos \theta_R = bA^{1/3} \cos^{1/3}\theta_a \cos(\theta_b + \theta_a/3) \quad (286)$$

Equation (286) can be rewritten as

$$R_m = b'A^{1/3} \quad (287)$$

where for subactinide nuclei

$$b' = b \cos^{1/3}\theta_a \cos(\theta_b + \theta_a/3) \quad (288)$$

The effective radius constant b' decreases with increasing strength of the applied gravitational or electromagnetic field. For the relatively weak gravitational field of the earth the internal phase angles are very small and therefore for this case

$$b' \sim b \sim 1.2 \text{ fm} \quad (289)$$

For strong gravitational or electromagnetic fields b' is given by equation (288) with $b = 1.2 \text{ fm}$ and $b' < b$.

B. Binding Energy of Atomic Nuclei Located in an Electromagnetic or Gravitational Field.

The standard expression for the binding energy B of a nucleus (Z, A) is given by the liquid drop model as⁶⁰⁻⁶⁷

$$\begin{aligned} B &= E_v - E_s - E_c - E_{\text{sym}} + E_{\text{pair}} + E_{\text{shell}} \\ &= \alpha A - \gamma A^{2/3} - \delta Z^2/A^{1/3} - \beta(N - Z)^2/A + E_{\text{pair}} + E_{\text{shell}} \end{aligned} \quad (290)$$

where E_v , E_s , E_c , E_{sym} , E_{pair} and E_{shell} = volume, surface, Coulomb, symmetry, nuclear pairing and nuclear shell energies respectively, and where α , γ , δ and β = volume, surface, Coulomb and symmetry energy coefficients respectively, and where from equation (84) it follows that $N - Z = A - 2Z$. The average binding energy per nucleon $\epsilon = B/A$ is written as⁶⁰⁻⁶⁷

$$\begin{aligned}\epsilon &= \epsilon_v - \epsilon_s - \epsilon_c - \epsilon_{\text{sym}} + \epsilon_{\text{pair}} + \epsilon_{\text{shell}} \\ &= \alpha - \gamma/A^{1/3} - \delta Z^2/A^{4/3} - \beta[(N - Z)/A]^2 + \epsilon_{\text{pair}} + \epsilon_{\text{shell}}\end{aligned}\quad (291)$$

where ϵ_v , ϵ_s , ϵ_c , ϵ_{sym} , ϵ_{pair} and ϵ_{shell} = average volume energy per nucleon, average surface energy per nucleon, average Coulomb energy per nucleon, average symmetry energy per nucleon, average pairing energy per nucleon and the average shell energy per nucleon respectively. More complicated forms of the nuclear symmetry energy have been considered by including the effects of the nuclear bulk modulus.⁶⁸ However, in this paper only the simple Weizsäcker-Bethe form given in equation (291) is considered.

For a nucleus in the presence of an electromagnetic or gravitational field the complex number nuclear binding energy is written as

$$\begin{aligned}\bar{B} &= \bar{E}_v - \bar{E}_s - \bar{E}_c - \bar{E}_{\text{sym}} + \bar{E}_{\text{pair}} + \bar{E}_{\text{shell}} \\ &= \bar{\alpha}\bar{a} - \bar{\gamma}\bar{a}^{2/3} - \bar{\delta}\bar{z}^2/\bar{a}^{1/3} - \bar{\beta}(\bar{n} - \bar{z})^2/\bar{a} + \bar{E}_{\text{pair}} + \bar{E}_{\text{shell}}\end{aligned}\quad (292)$$

where \bar{z} , \bar{n} and \bar{a} are given by equations (40) through (42) respectively, \bar{E}_v , \bar{E}_s , \bar{E}_c , \bar{E}_{sym} , \bar{E}_{pair} and \bar{E}_{shell} = complex number volume, surface, Coulomb, symmetry, pairing and shell energies respectively, and where $\bar{\alpha}$, $\bar{\gamma}$, $\bar{\delta}$ and $\bar{\beta}$ = complex number volume surface, Coulomb and symmetry energy coefficients respectively. The mass formula coefficients are represented as

$$\bar{\alpha} = \alpha \exp(j\theta_\alpha) \quad \bar{\gamma} = \gamma \exp(j\theta_\gamma) \quad (293)$$

$$\bar{\delta} = \delta \exp(j\theta_\delta) \quad \bar{\beta} = \beta \exp(j\theta_\beta) \quad (294)$$

$$\bar{E}_{\text{pair}} = E_{\text{pair}} \exp(j\theta_{E_{\text{pair}}}) \quad \bar{E}_{\text{shell}} = E_{\text{shell}} \exp(j\theta_{E_{\text{shell}}}) \quad (295)$$

The average complex number binding energy per nucleon $\bar{\epsilon}$ is written as

$$\begin{aligned}\bar{\epsilon} &= \bar{E}_v - \bar{E}_s - \bar{E}_c - \bar{E}_{\text{sym}} + \bar{E}_{\text{pair}} + \bar{E}_{\text{shell}} \\ &= \bar{\alpha} - \bar{\gamma}/\bar{a}^{1/3} - \bar{\delta}\bar{z}^2/\bar{a}^{4/3} - \bar{\beta}[(\bar{n} - \bar{z})/\bar{a}]^2 + \bar{\epsilon}_{\text{pair}} + \bar{\epsilon}_{\text{shell}}\end{aligned}\quad (296)$$

where \bar{E}_v , \bar{E}_s , \bar{E}_c , \bar{E}_{sym} , \bar{E}_{pair} and \bar{E}_{shell} = complex number average volume, surface, Coulomb, symmetry, pairing and shell energies per nucleon respectively.

The complex number neutron excess that appears in the symmetry energy terms in equations (292) and (296) is written as

$$\bar{\xi} = \xi \exp(j\theta_{\xi}) = \bar{n} - \bar{z} = W + \bar{a} - 2\bar{z} \quad (297)$$

where \bar{n} , \bar{z} and \bar{a} are given by equations (40) through (42) and are related by equation (82). Then for subactinide nuclei

$$\begin{aligned} \xi^2 &= n^2 + z^2 - 2zn \cos(\theta_n - \theta_z) \\ &= N^2 \cos^2 \theta_n + Z^2 \cos^2 \theta_z - 2ZN \cos \theta_z \cos \theta_n \cos(\theta_n - \theta_z) \end{aligned} \quad (298)$$

$$\begin{aligned} \tan \theta_{\xi} &= (n \sin \theta_n - z \sin \theta_z) / (n \cos \theta_n - z \cos \theta_z) \\ &= (N \cos \theta_n \sin \theta_n - Z \cos \theta_z \sin \theta_z) / (N \cos^2 \theta_n - Z \cos^2 \theta_z) \end{aligned} \quad (299)$$

For the approximation $\theta_z \sim \theta_n$, which is valid near the valley of beta stability, it follows from equations (298) and (299) that for subactinide nuclei

$$\xi \sim n - z = N \cos \theta_n - Z \cos \theta_z \sim (N - Z) \cos \theta_z \quad (300)$$

$$\theta_{\xi} \sim \theta_z \sim \theta_n \quad (301)$$

Combining equations (296) and (297) gives

$$\bar{\epsilon} = \bar{\alpha} - \bar{\gamma}/\bar{a}^{1/3} - \bar{\delta}\bar{z}^2/\bar{a}^{4/3} - \bar{\beta}(\bar{\xi}/\bar{a})^2 + \bar{\epsilon}_{\text{pair}} + \bar{\epsilon}_{\text{shell}} \quad (302)$$

From equation (297) it follows that the following approximation is valid for $W = 0$ in the subactinide elements

$$\begin{aligned} \bar{\xi}/\bar{a} &= \xi/a \exp[j(\theta_{\xi} - \theta_a)] \\ &\sim 1 - 2\bar{z}/\bar{a} \\ &\sim 1 - 2(Z \cos \theta_z) / (A \cos \theta_a) \exp[j(\theta_z - \theta_a)] \end{aligned} \quad (303)$$

Combining equations (302) and (303) gives the following approximation

$$\bar{\epsilon} = \bar{\alpha} - \bar{\gamma}/\bar{a}^{1/3} - \bar{\delta}\bar{z}^2/\bar{a}^{4/3} - \bar{\beta}(1 - 2\bar{z}/\bar{a})^2 + \bar{\epsilon}_{\text{pair}} + \bar{\epsilon}_{\text{shell}} \quad (304)$$

which is an approximate form of the exact equation (302).

Equation (292) can be written for the subactinides as

$$\begin{aligned} \bar{B} &= \alpha A \cos \theta_a \exp[j(\theta_{\alpha} + \theta_a)] - \gamma A^{2/3} \cos^{2/3} \theta_a \exp[j(\theta_{\gamma} + 2/3\theta_a)] \\ &\quad - \delta Z^2 A^{-1/3} \cos^2 \theta_z \cos^{-1/3} \theta_a \exp[j(\theta_{\delta} + 2\theta_z - 1/3\theta_a)] \\ &\quad - \beta \xi^2 A^{-1} \cos^{-1} \theta_a \exp[j(\theta_{\beta} + 2\theta_{\xi} - \theta_a)] \\ &\quad + E_{\text{pair}} \exp(j\theta_{E_{\text{pair}}}) + E_{\text{shell}} \exp(j\theta_{E_{\text{shell}}}) \end{aligned} \quad (305)$$

As an approximation the phase angles of the terms in equation (305) are taken to be equal

$$\begin{aligned}\theta_B &\sim \theta_\alpha + \theta_a \sim \theta_\gamma + 2/3\theta_a \\ &\sim \theta_\delta + 2\theta_z - 1/3\theta_a \sim \theta_\beta + 2\theta_\xi - \theta_a \\ &\sim \theta_{E_{\text{pair}}} \sim \theta_{E_{\text{shell}}}\end{aligned}\quad (306)$$

Then the magnitude of the binding energy is obtained from equation (305) to be

$$\begin{aligned}B &\sim \alpha A \cos \theta_a - \gamma A^{2/3} \cos^{2/3} \theta_a - \delta Z^2 A^{-1/3} \cos^2 \theta_z \cos^{-1/3} \theta_a \\ &\quad - \beta \xi^2 A^{-1} \cos^{-1} \theta_a + E_{\text{pair}} + E_{\text{shell}}\end{aligned}\quad (307)$$

Equation (88) gives $\theta_a = \theta_a(\theta_z, \theta_n, Z, A)$ so that in all further calculations it should be understood that θ_a is not really an independent variable. The approximation $\theta_z \sim \theta_n$ allows equation (307) to be written as

$$\begin{aligned}B &\sim \alpha A \cos \theta_a - \gamma A^{2/3} \cos^{2/3} \theta_a - \delta Z^2 A^{-1/3} \cos^2 \theta_z \cos^{-1/3} \theta_a \\ &\quad - \beta (N - Z)^2 A^{-1} \cos^2 \theta_z \cos^{-1} \theta_a + E_{\text{pair}} + E_{\text{shell}}\end{aligned}\quad (308)$$

The further approximation $\theta_\xi \sim \theta_z \sim \theta_n \sim \theta_a$ allows equation (306) and (307) to be written for the subactinides as

$$\begin{aligned}\theta_B &\sim \theta_\alpha + \theta_a \sim \theta_\gamma + 2/3\theta_a \\ &\sim \theta_\delta + 5/3\theta_a \sim \theta_\beta + \theta_a \\ &\sim \theta_{E_{\text{pair}}} \sim \theta_{E_{\text{shell}}}\end{aligned}\quad (309)$$

$$\begin{aligned}B &\sim \alpha A \cos \theta_a - \gamma A^{2/3} \cos^{2/3} \theta_a - \delta Z^2 A^{-1/3} \cos^{5/3} \theta_a \\ &\quad - \beta (N - Z)^2 A^{-1} \cos \theta_a + E_{\text{pair}} + E_{\text{shell}}\end{aligned}\quad (310)$$

which are valid in the valley of beta stability.

Each term in the expression for the binding energy and average binding energy per nucleon will now be considered separately with the exception of the nuclear pairing and shell effects which are more complicated and are not considered in this paper.

a. Volume Energy Term for Subactinide Nuclei.

The volume energy terms are written as

$$\begin{aligned}\bar{E}_v &= \bar{\alpha}\bar{a} = \alpha a \exp[j(\theta_\alpha + \theta_a)] \\ &= \alpha A \cos \theta_a \exp[j(\theta_\alpha + \theta_a)]\end{aligned}\quad (311)$$

$$\bar{e}_v = \bar{a} = \alpha \exp(j\theta_\alpha) \quad (311A)$$

The volume energy per nucleon \bar{a} describes the energy per nucleon of infinite nuclear matter but evaluated at the central density of a nucleus.⁶⁰⁻⁶⁸ The following complex number generalization of a simple density dependent expression for \bar{a} can be used⁶⁸

$$\bar{a} = \bar{b}_2 \bar{k}_c^2 - \bar{b}_3 \bar{k}_c^3 + \bar{b}_5 \bar{k}_c^5 + \bar{a}_{ex} \quad (312)$$

where

$$\bar{b}_2 = b_2 \exp(j\theta_{b2}) \quad \bar{b}_3 = b_3 \exp(j\theta_{b3}) \quad (313)$$

$$\bar{b}_5 = b_5 \exp(j\theta_{b5}) \quad \bar{a}_{ex} = \alpha_{ex} \exp(j\theta_{aex}) \quad (314)$$

where \bar{a}_{ex} = complex number exchange energy term, and where \bar{k}_c = complex number fermi wave number which is related to the complex number central density of a nucleus. The complex particle number density at the center of a nucleus is given by the following generalization of the standard scalar result⁵⁶

$$\bar{n}_c = 2/(3\pi^2) \bar{k}_c^3 \quad (315)$$

where

$$\bar{n}_c = n_c \exp(j\theta_{nc}) \quad \bar{k}_c = k_c \exp(j\theta_{kc}) \quad (316)$$

$$n_c = 2/(3\pi^2) k_c^3 \quad \theta_{nc} = 3\theta_{kc} \quad (317)$$

where $k_c \sim 1.35 \text{ fm}^{-1}$. The values of α and θ_α can be obtained by taking the real and imaginary parts of equation (312) as follows

$$\begin{aligned}\alpha \cos \theta_\alpha &= b_2 k_c^2 \cos(\theta_{b2} + 2\theta_{kc}) - b_3 k_c^3 \cos(\theta_{b3} + 3\theta_{kc}) \\ &\quad + b_5 k_c^5 \cos(\theta_{b5} + 5\theta_{kc}) + \alpha_{ex} \cos \theta_{aex}\end{aligned}\quad (318)$$

$$\begin{aligned}\alpha \sin \theta_\alpha &= b_2 k_c^2 \sin(\theta_{b2} + 2\theta_{kc}) - b_3 k_c^3 \sin(\theta_{b3} + 3\theta_{kc}) \\ &\quad + b_5 k_c^5 \sin(\theta_{b5} + 5\theta_{kc}) + \alpha_{ex} \sin \theta_{aex}\end{aligned}\quad (319)$$

In general $\theta_{b2} = 0$.

b. Surface Energy Term for Subactinide Nuclei.

The surface energy terms are written as

$$\bar{E}_s = \bar{\gamma} \bar{a}^{-2/3} = \gamma a^{2/3} \exp[j(\theta_\gamma + 2/3\theta_a)] \quad (320)$$

$$= \gamma A^{2/3} \cos^{2/3} \theta_a \exp[j(\theta_\gamma + 2/3\theta_a)]$$

$$\bar{\epsilon}_s = \bar{\gamma} \bar{a}^{-1/3} = \gamma a^{-1/3} \exp[j(\theta_\gamma - 1/3\theta_a)] \quad (321)$$

$$= \gamma A^{-1/3} \cos^{-1/3} \theta_a \exp[j(\theta_\gamma - 1/3\theta_a)]$$

The complex number surface energy coefficient can be written as the following generalization of the scalar result⁶⁸

$$\bar{\gamma} = \bar{c}_2 \bar{k}_c^2 - \bar{c}_3 \bar{k}_c^3 + \bar{c}_5 \bar{k}_c^5 + \bar{\gamma}_{ex} \quad (322)$$

where

$$\bar{c}_2 = c_2 \exp(j\theta_{c2}) \quad \bar{c}_3 = c_3 \exp(j\theta_{c3}) \quad (323)$$

$$\bar{c}_5 = c_5 \exp(j\theta_{c5}) \quad \bar{\gamma}_{ex} = \gamma_{ex} \exp(j\theta_{\gamma ex}) \quad (324)$$

where \bar{k}_c is related to the complex particle number density at the center of a nucleus by equation (315), and where $\bar{\gamma}_{ex}$ = exchange energy term for the nuclear surface. Formally one can calculate γ and θ_γ by taking the real and imaginary parts of equation (322) as follows

$$\gamma \cos \theta_\gamma = c_2 k_c^2 \cos(\theta_{c2} + 2\theta_{kc}) - c_3 k_c^3 \cos(\theta_{c3} + 3\theta_{kc}) \quad (325)$$

$$+ c_5 k_c^5 \cos(\theta_{c5} + 5\theta_{kc}) + \gamma_{ex} \cos \theta_{\gamma ex}$$

$$\gamma \sin \theta_\gamma = c_2 k_c^2 \sin(\theta_{c2} + 2\theta_{kc}) - c_3 k_c^3 \sin(\theta_{c3} + 3\theta_{kc}) \quad (326)$$

$$+ c_5 k_c^5 \sin(\theta_{c5} + 5\theta_{kc}) + \gamma_{ex} \sin \theta_{\gamma ex}$$

from which γ and θ_γ are easily obtained.

c. Coulomb Energy Term for Subactinide Nuclei.

The complex number Coulomb energy terms are written as

$$\bar{E}_c = \bar{\delta z}^2 / \bar{a}^{1/3} = \delta z^2 a^{-1/3} \exp[j(\theta_\delta + 2\theta_z - 1/3\theta_a)] \quad (327)$$

$$= \delta z^2 A^{-1/3} \cos^2 \theta_z \cos^{-1/3} \theta_a \exp[j(\theta_\delta + 2\theta_z - 1/3\theta_a)]$$

$$\begin{aligned}\bar{\epsilon}_c &= \bar{\delta} \bar{z}^2 / \bar{a}^{4/3} = \delta z^2 a^{-4/3} \exp[j(\theta_\delta + 2\theta_z - 4/3\theta_a)] \\ &= \delta Z^2 A^{-4/3} \cos^2 \theta_z \cos^{-4/3} \theta_a \exp[j(\theta_\delta + 2\theta_z - 4/3\theta_a)]\end{aligned}\quad (328)$$

The complex number Coulomb energy coefficient can be written as a generalization of the scalar result⁶⁸

$$\begin{aligned}\bar{\delta} &= 3/5(e^2/\bar{b}) = 0.863/\bar{b} \quad \text{MeV} \\ &= (0.863/1.523)\bar{k}_c \quad \text{MeV}\end{aligned}\quad (329)$$

where \bar{b} = complex number radius parameter defined in equation (283). Equation (329) is equivalent to

$$\delta = 0.863/b = (0.863/1.523)k_c \quad \text{MeV} \quad (330)$$

$$\theta_\delta = -\theta_b = \theta_{kc} \quad (331)$$

For simplicity it will be assumed that the wave number k_c of the central density of an atomic nucleus is approximately equal to the wave number of infinite nuclear matter k_F , so that

$$k_c \sim k_F = 1.35 \text{ fm}^{-1} \quad (332)$$

but in fact k_c is slightly larger or smaller than k_F due to Coulomb and surface forces.⁶⁸

d. Symmetry Energy Term for Subactinide Nuclei.

The complex number symmetry energy terms are written as

$$\begin{aligned}\bar{\epsilon}_{\text{sym}} &= \bar{\beta} \bar{\xi}^2 / \bar{a} = \beta \xi^2 a^{-1} \exp[j(\theta_\beta + 2\theta_\xi - \theta_a)] \\ &= \beta \xi^2 A^{-1} \cos^{-1} \theta_a \exp[j(\theta_\beta + 2\theta_\xi - \theta_a)]\end{aligned}\quad (333)$$

$$\begin{aligned}\bar{\epsilon}_{\text{sym}} &= \bar{\beta} (\bar{\xi}/\bar{a})^2 = \beta \xi^2 a^{-2} \exp[j(\theta_\beta + 2\theta_\xi - 2\theta_a)] \\ &= \beta \xi^2 A^{-2} \cos^{-2} \theta_a \exp[j(\theta_\beta + 2\theta_\xi - 2\theta_a)]\end{aligned}\quad (334)$$

where $\bar{\xi}$, ξ and θ_ξ are given by equations (297) through (299) respectively. Equations (333) and (334) can be simplified by assuming the approximation $\theta_z \sim \theta_n$, then equations (298) and (299) give

$$\xi \sim (N - Z) \cos \theta_z \quad \theta_\xi \sim \theta_z \sim \theta_n \quad (335)$$

and equations (333) and (334) become

$$\bar{E}_{\text{sym}} \sim \beta(N-Z)^2 A^{-1} \cos^{-1} \theta_a \cos^2 \theta_z \exp[j(\theta_\beta + 2\theta_z - \theta_a)] \quad (336)$$

$$\bar{e}_{\text{sym}} \sim \beta(N-Z)^2 A^{-2} \cos^{-2} \theta_a \cos^2 \theta_z \exp[j(\theta_\beta + 2\theta_z - 2\theta_a)] \quad (337)$$

As a further approximation let $\theta_z \sim \theta_n \sim \theta_a$, which follows from equation (88) for small values of $\theta_z \sim \theta_n$, then equations (336) and (337) become

$$\bar{E}_{\text{sym}} \sim \beta(N-Z)^2 A^{-1} \cos \theta_a \exp[j(\theta_\beta + \theta_a)] \quad (338)$$

$$\bar{e}_{\text{sym}} \sim \beta(N-Z)^2 A^{-2} \exp(j\theta_\beta) \quad (339)$$

Equations (338) and (339) are valid in the vicinity of the valley of beta stability where $\theta_z \sim \theta_n \sim \theta_a$.

The complex number symmetry energy coefficient can be written as a generalization of a corresponding scalar form as follows⁶⁸

$$\bar{\beta} = \bar{e}_2 \bar{k}_c^2 - \bar{e}_3 \bar{k}_c^3 + \bar{e}_5 \bar{k}_c^5 + \bar{\beta}_{\text{ex}} \quad (340)$$

where

$$\bar{e}_2 = e_2 \exp(j\theta_{e2}) \quad \bar{e}_3 = e_3 \exp(j\theta_{e3}) \quad (341)$$

$$\bar{e}_5 = e_5 \exp(j\theta_{e5}) \quad \bar{\beta}_{\text{ex}} = \beta_{\text{ex}} \exp(j\theta_{\beta\text{ex}}) \quad (342)$$

where $\bar{\beta}_{\text{ex}}$ = complex number exchange energy contribution to the quadratic symmetry energy coefficient. The values of β and θ_β can be determined from the real and imaginary components of equation (340) which are given by

$$\begin{aligned} \beta \cos \theta_\beta &= e_2 k_c^2 \cos(\theta_{e2} + 2\theta_{kc}) - e_3 k_c^3 \cos(\theta_{e3} + 3\theta_{kc}) \\ &\quad + e_5 k_c^5 \cos(\theta_{e5} + 5\theta_{kc}) + \beta_{\text{ex}} \cos \theta_{\beta\text{ex}} \end{aligned} \quad (343)$$

$$\begin{aligned} \beta \sin \theta_\beta &= e_2 k_c^2 \sin(\theta_{e2} + 2\theta_{kc}) - e_3 k_c^3 \sin(\theta_{e3} + 3\theta_{kc}) \\ &\quad + e_5 k_c^5 \sin(\theta_{e5} + 5\theta_{kc}) + \beta_{\text{ex}} \sin \theta_{\beta\text{ex}} \end{aligned} \quad (344)$$

In general $\theta_{e2} = 0$ because this term arises from the kinetic energy of a non-interacting Fermi gas, however $\theta_{e3} \neq 0$ and $\theta_{e5} \neq 0$ because the cubic and fifth order terms arise from nuclear potentials that are complex numbers. In more complicated forms of the nuclear mass formula the symmetry energy is separated into volume and surface components.⁶⁸

C. Measured Binding Energies of Subactinide Nuclei Located in an External Field.

The real and imaginary parts of equation (292) are given by

$$\begin{aligned}
B \cos \theta_B = & \alpha A \cos \theta_a \cos(\theta_\alpha + \theta_a) - \gamma A^{2/3} \cos^{2/3} \theta_a \cos(\theta_\gamma + 2/3 \theta_a) \\
& - \delta Z^2 A^{-1/3} \cos^2 \theta_z \cos^{-1/3} \theta_a \cos(\theta_\delta + 2\theta_z - 1/3 \theta_a) \\
& - \beta \xi^2 A^{-1} \cos^{-1} \theta_a \cos(\theta_\beta + 2\theta_z - \theta_a) \\
& + E_{\text{pair}} \cos \theta_{\text{Epair}} + E_{\text{shell}} \cos \theta_{\text{Eshell}}
\end{aligned} \quad (345)$$

$$\begin{aligned}
B \sin \theta_B = & \alpha A \cos \theta_a \sin(\theta_\alpha + \theta_a) - \gamma A^{2/3} \cos^{2/3} \theta_a \sin(\theta_\gamma + 2/3 \theta_a) \\
& - \delta Z^2 A^{-1/3} \cos^2 \theta_z \cos^{-1/3} \theta_a \sin(\theta_\delta + 2\theta_z - 1/3 \theta_a) \\
& - \beta \xi^2 A^{-1} \cos^{-1} \theta_a \sin(\theta_\beta + 2\theta_z - \theta_a) \\
& + E_{\text{pair}} \sin \theta_{\text{Epair}} + E_{\text{shell}} \sin \theta_{\text{Eshell}}
\end{aligned} \quad (346)$$

Equations (345) and (346) immediately determine B and θ_B . The measured binding energy is just the real part of the complex number binding energy, so that

$$B_m = \alpha_m A - \gamma_m A^{2/3} - \delta_m Z^2 A^{-1/3} - \beta'_m \xi^2 A^{-1} + E_{\text{pair}}^m + E_{\text{shell}}^m \quad (347)$$

where for subactinide nuclei

$$\alpha_m = \alpha \cos \theta_a \cos(\theta_\alpha + \theta_a) \quad (348)$$

$$\gamma_m = \gamma \cos^{2/3} \theta_a \cos(\theta_\gamma + 2/3 \theta_a) \quad (349)$$

$$\delta_m = \delta \cos^2 \theta_z \cos^{-1/3} \theta_a \cos(\theta_\delta + 2\theta_z - 1/3 \theta_a) \quad (350)$$

$$\beta'_m = \beta \cos^{-1} \theta_a \cos(\theta_\beta + 2\theta_z - \theta_a) \quad (351)$$

where $\xi = \xi(Z, N, \theta_z, \theta_n)$ and is defined by equation (298), and $\theta_\xi = \theta_\xi(Z, N, \theta_z, \theta_n)$ is defined by equation (299). The internal phase angle θ_a is given by equation (88) to be $\theta_a = \theta_a(Z, N, \theta_z, \theta_n)$.

If the approximation $\theta_\xi \sim \theta_z \sim \theta_n$ is assumed, then the approximations in equations (335) and (336) allow equation (347) to be written as

$$B_m = \alpha_m A - \gamma_m A^{2/3} - \delta_m Z^2 A^{-1/3} - \beta_m (N - Z)^2 A^{-1} + E_{\text{pair}}^m + E_{\text{shell}}^m \quad (352)$$

where now

$$\beta_m = \beta \cos^2 \theta_z \cos^{-1} \theta_a \cos(\theta_\beta + 2\theta_z - \theta_a) \quad (353)$$

If the further approximation $\theta_\xi \sim \theta_z \sim \theta_a$ is made, then equation (352) is the measured binding energy with

$$\alpha_m = \alpha \cos \theta_a \cos(\theta_\alpha + \theta_a) \quad (354)$$

$$\gamma_m = \gamma \cos^{2/3} \theta_a \cos(\theta_\gamma + 2/3\theta_a) \quad (355)$$

$$\delta_m = \delta \cos^{5/3} \theta_a \cos(\theta_\delta + 5/3\theta_a) \quad (356)$$

$$\beta_m = \beta \cos \theta_a \cos(\theta_\beta + \theta_a) \quad (357)$$

which are useful for nuclei near the valley of beta stability. The measured values of the symmetry energy coefficients are^{60,61}

$$\alpha_m = 15.5 \quad \gamma_m = 17.2 \quad \delta_m = 0.698 \quad \beta_m = 23.3 \text{ MeV} \quad (358)$$

Equations (354) through (357) show that

$$\alpha_m < \alpha \quad \gamma_m < \gamma \quad \delta_m < \delta \quad \beta_m < \beta \quad (359)$$

The values of the nuclear mass formula parameters θ_z , θ_n , θ_a , α , θ_α , γ , θ_γ , δ , θ_δ , β and θ_β can be obtained by fitting the real part of the complex number binding energy given by equation (347) to the measured values of the atomic masses of the elements. A simplified procedure uses the approximation $\theta_z \sim \theta_n$ and equation (352) for the fit to atomic masses. Expressions for the atomic masses of the elements will now be considered.

D. Masses of Atoms Located in an Electromagnetic or Gravitational Field.

The conventional relationship between atomic mass and nuclear binding energy is written as⁶¹

$$M = Zm_H + Nm_n - B \quad (360)$$

where M = atomic mass of an element, m_H = mass of hydrogen atom and m_n = neutron mass. In the presence of an external field the atomic mass is a complex number in an internal space and is given by

$$\bar{M} = \bar{Z}m_H + \bar{N}m_n - \bar{B} \quad (361)$$

where the complex number atomic mass is written as

$$\bar{M} = M \exp(j\theta_M) \quad (362)$$

Using equations (76), (77), (79) and (80) allows the real and imaginary parts of equation (361) to be written as

$$M \cos \theta_M = G \quad (363)$$

$$M \sin \theta_M = F \quad (364)$$

where for subactinide nuclei

$$F = m_H Z \cos \theta_z \sin \theta_z + m_n N \cos \theta_n \sin \theta_n - B \sin \theta_B \quad (365)$$

$$G = m_H Z \cos^2 \theta_z + m_n N \cos^2 \theta_n - B \cos \theta_B \quad (366)$$

Equations (363) through (366) can be used to obtain M and θ_M as

$$\tan \theta_M = F/G \quad (367)$$

$$M^2 = F^2 + G^2 \quad (368)$$

$$\begin{aligned} &= B^2 + m_H^2 Z^2 \cos^2 \theta_z + m_n^2 N^2 \cos^2 \theta_n \\ &\quad + 2m_H m_n ZN \cos \theta_z \cos \theta_n \cos(\theta_z - \theta_n) \\ &\quad - 2m_H BZ \cos \theta_z \cos(\theta_z - \theta_B) \\ &\quad - 2m_n BN \cos \theta_n \cos(\theta_n - \theta_B) \end{aligned}$$

The measured atomic mass is given by equation (363) which can be rewritten as

$$M_m = m_H Z \cos^2 \theta_z + m_n N \cos^2 \theta_n - B_m \quad (369)$$

The approximation $\theta_z \sim \theta_n \sim \theta_B$ combined with equations (361) through (369) gives

$$\theta_M \sim \theta_z \sim \theta_n \sim \theta_B \quad (370)$$

$$M \sim (m_H Z + m_n N) \cos \theta_z - B \quad (371)$$

$$M_m \sim (m_H Z + m_n N) \cos^2 \theta_z - B_m \quad (372)$$

where θ_a is given by equations (88) and (89A) within this approximation. Note that M, M_m and θ_M vary with the strength of the applied electromagnetic field because $\theta_z = \theta_z(H)$ in equations (370) through (372), but the following intrinsic mass is a constant independent of the applied electromagnetic (or gravitational) field

$$m_H Z + m_n N = \text{constant} \quad (373)$$

and represents the universal law of the conservation of rest mass and baryon number. For the case $\theta_z \sim \theta_n \sim \theta_a \sim \theta_B$ the mass relations in equations (370) through (372) become for subactinide nuclei

$$\theta_M \sim \theta_a \quad (374)$$

$$M \sim (m_H Z + m_N) \cos \theta_a - B(Z, N, \theta_a) \quad (375)$$

$$M_m \sim (m_H Z + m_N) \cos^2 \theta_a - B_m(Z, N, \theta_a) \quad (376)$$

which are valid near the valley of beta stability of subactinide nuclei.

The variation of the measured atomic mass and the magnitude of the atomic mass with the strength of the external electromagnetic field can be obtained from equations (368) and (369) by the following formulas

$$dM_m/dH = \partial M_m / \partial \theta_z d\theta_z/dH + \partial M_m / \partial \theta_n d\theta_n/dH + \partial M_m / \partial \theta_a d\theta_a/dH \quad (377)$$

$$dM/dH = \partial M / \partial \theta_z d\theta_z/dH + \partial M / \partial \theta_n d\theta_n/dH + \partial M / \partial \theta_a d\theta_a/dH \quad (378)$$

The internal phase angle $\theta_a = \theta_a(\theta_z, \theta_n, Z, N)$ is given by equation (88) so that the derivative $d\theta_a/dH$ can be evaluated as

$$d\theta_a/dH = \partial \theta_a / \partial \theta_z d\theta_z/dH + \partial \theta_a / \partial \theta_n d\theta_n/dH \quad (379)$$

Therefore equations (377) and (378) can be written as

$$\begin{aligned} dM_m/dH &= (\partial M_m / \partial \theta_z + \partial M_m / \partial \theta_a \partial \theta_a / \partial \theta_z) d\theta_z/dH \\ &+ (\partial M_m / \partial \theta_n + \partial M_m / \partial \theta_a \partial \theta_a / \partial \theta_n) d\theta_n/dH \end{aligned} \quad (380)$$

$$\begin{aligned} dM/dH &= (\partial M / \partial \theta_z + \partial M / \partial \theta_a \partial \theta_a / \partial \theta_z) d\theta_z/dH \\ &+ (\partial M / \partial \theta_n + \partial M / \partial \theta_a \partial \theta_a / \partial \theta_n) d\theta_n/dH \end{aligned} \quad (381)$$

where for example equation (369) gives for subactinide nuclei

$$\partial M_m / \partial \theta_z = -2m_H Z \cos \theta_z \sin \theta_z - \partial B_m / \partial \theta_z \quad (382)$$

$$\partial M_m / \partial \theta_n = -2m_N N \cos \theta_n \sin \theta_n - \partial B_m / \partial \theta_n \quad (383)$$

$$\partial M_m / \partial \theta_a = -\partial B_m / \partial \theta_a \quad (384)$$

where B_m is given by equation (347).

For the approximate case $\theta_z \sim \theta_n \sim \theta_a$, which follows from equation (88) for small arguments, it follows that

$$dM_m/dH = dM_m/d\theta_a d\theta_a/dH \quad (385)$$

$$dM/dH = dM/d\theta_a d\theta_a/dH \quad (386)$$

where equations (375) and (376) give for $\theta_z \sim \theta_n \sim \theta_a$ and for subactinide nuclei

$$dM_m/d\theta_a = -2(m_H Z + m_N) \cos \theta_a \sin \theta_a - dB_m/d\theta_a \quad (387)$$

$$dM/d\theta_a = - (m_H Z + m_N) \sin \theta_a - dB/d\theta_a \quad (388)$$

where the approximate value of B_m given by equations (352) and (354) through (357) are used in conjunction with equation (387), while the approximate value of B given by equation (310) is used in conjunction with equation (388). Therefore from equation (352) and within the approximation $\theta_z \sim \theta_n \sim \theta_a$, the derivative in equation (387) is given by

$$\begin{aligned} dB_m/d\theta_a = & d\alpha_m/d\theta_a A - d\gamma_m/d\theta_a A^{2/3} - d\delta_m/d\theta_a Z^2 A^{-1/3} \\ & - d\beta_m/d\theta_a (N - Z)^2 A^{-1} + dE_{\text{pair}}^m/d\theta_a + dE_{\text{shell}}^m/d\theta_a \end{aligned} \quad (389)$$

where from equations (354) through (357) for $\theta_z \sim \theta_n \sim \theta_a$ in subactinide nuclei

$$d\alpha_m/d\theta_a = -\alpha[\sin \theta_a \cos(\theta_\alpha + \theta_a) + \cos \theta_a \sin(\theta_\alpha + \theta_a)] \quad (390)$$

$$d\gamma_m/d\theta_a = -2/3\gamma[\cos^{-1/3}\theta_a \sin \theta_a \cos(\theta_\gamma + 2/3\theta_a) + \cos^{2/3}\theta_a \sin(\theta_\gamma + 2/3\theta_a)] \quad (391)$$

$$d\delta_m/d\theta_a = -5/3\delta[\cos^{2/3}\theta_a \sin \theta_a \cos(\theta_\delta + 2/3\theta_a) + \cos^{5/3}\theta_a \sin(\theta_\delta + 5/3\theta_a)] \quad (392)$$

$$d\beta_m/d\theta_a = -\beta[\sin \theta_a \cos(\theta_\beta + \theta_a) + \cos \theta_a \sin(\theta_\beta + \theta_a)] \quad (393)$$

For the approximation $\theta_z = \theta_n = \theta_a$, the derivative of the magnitude of the binding energy that appears in equation (388) is obtained from equation (310) to be

$$\begin{aligned} dB/d\theta_a = & \sin \theta_a [-\alpha A + 2/3\gamma A^{2/3} \cos^{-1/3}\theta_a + 5/3\delta Z^2 A^{-1/3} \cos^{2/3}\theta_a \\ & + \beta(N - Z)^2 A^{-1}] + dE_{\text{pair}}/d\theta_a + dE_{\text{shell}}/d\theta_a \end{aligned} \quad (394)$$

From equations (387) through (394) it follows that for subactinide nuclei

$$dB_m/d\theta_a < 0 \quad dB/d\theta_a < 0 \quad (395)$$

$$dM_m/d\theta_a < 0 \quad dM/d\theta_a < 0 \quad (396)$$

For example, within the approximation $\theta_z \sim \theta_n \sim \theta_a$ these equations give

$$\begin{aligned} dM/d\theta_a = & -\sin \theta_a [m_H Z + m_N - \alpha A + 2/3\gamma A^{2/3} \cos^{-1/3}\theta_a \\ & + 5/3\delta Z^2 A^{-1/3} \cos^{2/3}\theta_a + \beta(N - Z)^2 A^{-1}] \\ & - dE_{\text{pair}}/d\theta_a - dE_{\text{shell}}/d\theta_a \end{aligned} \quad (397)$$

which is a negative number because of the dominant contribution of the rest mass terms.

E. Valley of Beta Stability for Nuclei with Broken Internal Symmetries.

Radioactive beta decays require that the complex atomic number \bar{z} adjust itself so as to minimize the binding energy of a nucleus given by equation (292) but subject to the constraints represented in equations (82) and (84).⁶⁶ Combining equations (82) and (292) gives after neglecting shell and pairing energy effects

$$\bar{B} = \bar{\alpha}\bar{a} - \bar{\gamma}\bar{a}^{2/3} - \bar{\delta}\bar{z}^2/\bar{a}^{1/3} - \bar{\beta}(\bar{a} - 2\bar{z} + W)^2/\bar{a} \quad (398)$$

The minimum binding energy condition is

$$\partial\bar{B}/\partial\bar{z}|_{\bar{a}} = -2\bar{\delta}\bar{z}/\bar{a}^{1/3} + 4\bar{\beta}(\bar{a} - 2\bar{z} + W)/\bar{a} = 0 \quad (399)$$

which gives using equation (82) with $W = 0$

$$\bar{z}^{vs} = \bar{a}/2(1 + \bar{c}\bar{a}^{2/3})^{-1} \quad (400)$$

$$\bar{n}^{vs} = \bar{a}/2(1 + 2\bar{c}\bar{a}^{2/3})(1 + \bar{c}\bar{a}^{2/3})^{-1} \quad (401)$$

where vs = valley of beta stability, and where

$$\bar{c} = \bar{\delta}/(4\bar{\beta}) \quad c = \delta/(4\beta) \quad \theta_c = \theta_\delta - \theta_\beta \quad (402)$$

For relatively small nuclei the beta stability condition is

$$\bar{z}^{vs} = \bar{a}/2 \quad z^{vs} = a/2 \quad \theta_z^{vs} = \theta_a \quad (403)$$

$$\bar{n}^{vs} = \bar{a}/2 \quad n^{vs} = a/2 \quad \theta_n^{vs} = \theta_a \quad (404)$$

Combining equations (40), (41) and (404) gives the beta stability condition for light nuclei as

$$Z^{vs} = A/2 \quad N^{vs} = A/2 \quad (405)$$

which is the standard beta stability condition for small nuclei.⁶⁵ Therefore for light nuclei, the standard scalar theory given by equation (405) is valid because $\theta_z^{vs} = \theta_a$. For medium weight nuclei the following approximation to equations (400) and (401) can be used

$$\bar{z}^{vs} = \bar{a}/2(1 - \bar{c}\bar{a}^{2/3}) \quad (406)$$

$$\bar{n}^{vs} = \bar{a}/2(1 + \bar{c}\bar{a}^{2/3}) \quad (407)$$

where \bar{c} is given by equation (402), and where $W = 0$.

For heavy nuclei the exact equation (400) must be used to calculate \bar{z}^{vs} .

Equation (400) can be written as

$$\bar{z}^{vs} = (G + jF)/D \quad (408)$$

$$\tan \theta_z^{vs} = F/G \quad z^{vs} = (G^2 + F^2)^{1/2}/D \quad (409)$$

where

$$G = a/2 \cos \theta_a [1 + ca^{2/3} \cos(\theta_c + 2/3\theta_a)] + c/2 a^{5/3} \sin \theta_a \sin(\theta_c + 2/3\theta_a) \quad (410)$$

$$F = a/2 \sin \theta_a [1 + ca^{2/3} \cos(\theta_c + 2/3\theta_a)] - c/2 a^{5/3} \cos \theta_a \sin(\theta_c + 2/3\theta_a) \quad (411)$$

$$D = [1 + ca^{2/3} \cos(\theta_c + 2/3\theta_a)]^2 + c^2 a^{4/3} \sin^2(\theta_c + 2/3\theta_a) \quad (412)$$

Combining equations (67) and (410) through (412) gives for subactinide nuclei

$$G = A/2 \cos^2 \theta_a [1 + cA^{2/3} \cos^{2/3} \theta_a \cos(\theta_c + 2/3\theta_a)] + c/2A^{5/3} \cos^{5/3} \theta_a \sin \theta_a \sin(\theta_c + 2/3\theta_a) \quad (413)$$

$$F = A/2 \cos \theta_a \sin \theta_a [1 + cA^{2/3} \cos^{2/3} \theta_a \cos(\theta_c + 2/3\theta_a)] - c/2A^{5/3} \cos^{8/3} \theta_a \sin(\theta_c + 2/3\theta_a) \quad (414)$$

$$D = [1 + cA^{2/3} \cos^{2/3} \theta_a \cos(\theta_c + 2/3\theta_a)]^2 + c^2 A^{4/3} \cos^{4/3} \theta_a \sin^2(\theta_c + 2/3\theta_a) \quad (415)$$

If A and θ_a are taken to be the known quantities, then equation (82) involves five unknown quantities W , Z , N , θ_z and θ_n . The complex number equation (82) and the scalar equation (84) supply three equations for determining the five unknown quantities. The complex number valley of beta stability equation (399) supplies two additional equations and so a complete solution is possible. From equations (409) and (413) through (415) that describe the valley of beta stability for the general case of nuclei of arbitrary size it follows that in general

$$\theta_z^{vs} = \theta_z^{vs}(A, \theta_a) \quad Z^{vs} = Z^{vs}(A, \theta_a) \quad (416)$$

Combining equations (67) and (409) gives for subactinide nuclei

$$Z^{vs} = z^{vs}/\cos \theta_z^{vs} = (G^2 + F^2)/(DG) \quad (417)$$

$$N^{vs} = A - Z^{vs} = A - (G^2 + F^2)/(DG) \quad (418)$$

The value of θ_n^{vs} is obtained exactly from equation (82) which is written in the form

$$\bar{n}^{vs} = \bar{a} + W^{vs} - \bar{z}^{vs} \quad (419)$$

but it is most easily obtained approximately by taking $W^{vs} = 0$ with the result

$$\tan \theta_n^{vs} = (A \cos \theta_a \sin \theta_a - Z^{vs} \cos \theta_z^{vs} \sin \theta_z^{vs}) (A \cos^2 \theta_a - Z^{vs} \cos^2 \theta_z^{vs})^{-1} \quad (420)$$

where θ_z^{vs} and Z^{vs} are given by equations (409). The measured values of the atomic number, neutron number and atomic mass number in the valley of beta stability are given for subactinide nuclei by

$$z_m^{vs} = Z^{vs} \cos^2 \theta_z^{vs} \quad n_m^{vs} = N^{vs} \cos^2 \theta_n^{vs} \quad a_m = A \cos^2 \theta_a \quad (421)$$

which are generally not integers.

Consider now the determination of the specific nucleus within the valley of beta stability that has the greatest average binding energy per nucleon when an external field is present. Combining equations (304) and (403) gives the average binding energy per nucleon for nuclei within the valley of beta stability approximately as

$$\bar{\epsilon}^{vs} = \bar{a} - \bar{\gamma}/\bar{a}^{1/3} - \bar{\delta}/4\bar{a}^{2/3} + \bar{\epsilon}_{pair}(\bar{a}) + \bar{\epsilon}_{shell}(\bar{a}) \quad (422)$$

Then the nucleus having the greatest average binding energy per nucleon after neglecting nuclear shell and pairing effects is given by

$$d\bar{\epsilon}^{vs}/d\bar{a} = 1/3\bar{\gamma}/\bar{a}^{4/3} - 1/6\bar{\delta}/\bar{a}^{1/3} = 0 \quad (423)$$

which gives approximately

$$\bar{a}^{gb} = 2\bar{\gamma}/\bar{\delta} \quad (424)$$

$$a^{gb} = 2\gamma/\delta \quad \theta_a^{gb} = \theta_\gamma - \theta_\delta \quad (425)$$

where gb = greatest binding. From equation (425) it follows that approximately

$$A^{gb} \cos \theta_a^{gb} = 2\gamma/\delta \quad (426)$$

or equivalently

$$A^{gb} = 2\gamma/\delta \sec(\theta_\gamma - \theta_\delta) \quad (427)$$

The conventional prediction of the most stable nucleus is⁵⁹⁻⁶⁶

$$A_c^{gb} = 2\gamma/\delta \quad (428)$$

and therefore for subactinide nuclei

$$A_c^{gb} = A_c^{gb} \sec(\theta_Y - \theta_\delta) > A_c^{gb} \quad (429 - 444)$$

The peak in the average binding energy per nucleon curve for nuclei in an external electromagnetic or gravitational field is shifted to a higher value of atomic mass number than is predicted by conventional calculations, but this effect is small if $\theta_Y \sim \theta_\delta$. The calculation of the nucleus of greatest average binding energy per nucleon within the valley of beta stability has been done using the approximation given in equations (403) and (404) for which case the nuclear symmetry energy term vanishes. An exact calculation requires the use of equations (400) and (401) for which the symmetry energy term in equation (304) does not vanish as it does in equation (422). The inclusion of the symmetry energy term in the calculation of the subactinide nucleus of greatest binding within the valley of beta stability is algebraically difficult and will not be done in this paper.

This section suggests that a nuclear mass formula which is suitable for subactinide nuclei in an external electromagnetic or gravitational field must include complex number values of the atomic number and atomic mass number and complex values of the volume, surface, Coulomb, symmetry, pairing and shell energy coefficients. Nuclear properties such as neutron binding energies and neutron capture and fission cross sections will also be complex numbers in an internal space, and will depend on the strength of the external field because the magnitudes and phase angles of all terms of a nuclear mass formula are affected by the external field. The measured nuclear properties will correspond to the calculated real values of the complex number nuclear properties. The particular forms chosen for the complex number representations of the atomic number, neutron number and atomic mass number refer generally speaking only to subactinide nuclei. More precisely, the forms chosen in equation (67) refer only to nuclei for which $\chi \leq 1$ where χ = fissility parameter, and this condition is generally valid for subactinide nuclei. A forthcoming paper will treat the case $\chi \geq 1$ which begins to occur in the actinides.

5. LOW ENERGY FISSION OF LIGHTER THAN ACTINIDE NUCLEI IN AN ELECTROMAGNETIC FIELD. This section derives the conditions required for the spontaneous or thermal neutron induced fission of lighter than actinide nuclei (subactinide nuclei) that are located in an electromagnetic field, determines the magnitude of the internal phase angle of the atomic number that is required to bring a subactinide nucleus into a state of incipient fission, and gives a simple relationship between this critical internal phase angle and the strength of the electromagnetic field required to catalyze the clean fission of subactinide nuclei. Under zero field conditions spontaneous or thermal neutron induced fission occurs only in some of the actinide and transactinide elements but does not occur in the subactinide elements.⁶⁹⁻⁸⁰ Some of these heavy elements such as ^{235}U and ^{239}Pu are used in conventional nuclear reactors. The possibility of thermal neutron induced fission in elements lighter than the actinides in the presence of an electromagnetic field will have practical applications to the design of clean fission nuclear reactors because the fission products will be relatively light elements which are either stable against beta decay or are only low level beta emitters. An atomic nucleus is a quantum many-body system which has both collective effects such as volume and surface energies, and independent

particle motion effects such as shell structure. Both collective motion and independent particle motion are important to the description of the fission process.⁶⁹⁻⁸⁰ For instance, the shell structure determines the nature of the doubly humped fission barrier.⁶⁹⁻⁸⁰ In this paper only the simple collective terms such as volume, surface, Coulomb and symmetry energies are considered for the description of the catalysis of thermal neutron induced fission in the sub-actinide nuclei by an applied electromagnetic field which in this case must be a γ ray field.

A. Bohr-Wheeler Fission Condition for Nuclei in an Electromagnetic Field.

The standard Bohr-Wheeler analysis for spontaneous or thermal neutron induced nuclear fission utilizes the fissility parameter which is defined by⁶⁹⁻⁸⁰

$$\chi = (Z^2/A)(\kappa\gamma/\delta)^{-1} \quad (445)$$

and the spontaneous and thermal neutron induced fission condition is written for actinide nuclei as⁶⁹⁻⁸⁰

$$\chi \geq 1 \quad Z^2/A \geq \kappa\gamma/\delta \quad (446)$$

where γ and δ = surface and Coulomb energy coefficients that appear in the liquid drop nuclear mass formula treated in Section 4, and where theoretically for spontaneous fission

$$\kappa = g/h = 2 \quad (447)$$

where $g = 2/5$ and $h = 1/5$ are the second order series expansion coefficients of the surface and Coulomb energies respectively when these terms are expanded in terms of an ellipsoidal deformation parameter.⁶⁹⁻⁸⁰ The values of κ , γ and δ along with the other mass formula parameters are determined empirically.⁶⁹⁻⁸⁰ The values of κ are different for spontaneous and for thermal neutron induced fission, and in fact κ is dependent on the energy of the incident neutrons.⁶⁹⁻⁸⁰ For thermal neutron induced fission⁶⁹⁻⁸⁰

$$\kappa \sim 1.471 \quad (448)$$

Choosing $\gamma = 17.2$ MeV and $\delta = 0.698$ MeV yields the following fission conditions for a zero value of the externally applied field⁶⁹⁻⁸⁰

$$Z^2/A > 49.28 \quad \text{spontaneous fission} \quad (449)$$

$$Z^2/A > 36.25 \quad \text{thermal neutron induced fission} \quad (450)$$

These inequalities show that, loosely speaking, only the actinides and trans-actinides can undergo spontaneous or thermal neutron induced fission, but not all of these heavy elements undergo fission. Within this group of heavy elements the more neutron rich isotopes tend to be more stable against fission, for example ^{238}U is stable against thermal neutron induced fission but ^{235}U is fissile. The empirical value of κ that describes thermal neutron induced fission will depend on the values selected for the mass formula parameters γ and δ . In general κ can be taken to be a decreasing function of the kinetic energy of

the incident neutrons. The fission criteria presented above ignore all shell structure effects and are therefore approximate relations which show only general behavior and for which counterexamples can always be found in the border region between fissile and non-fissile nuclei.

The generalization of equation (446) to the case of atomic nuclei located in an electromagnetic or gravitational field, which breaks the symmetry of the atomic number, neutron number and atomic mass number, can be written as

$$\bar{z}^2/\bar{a} > \kappa\bar{\gamma}/\bar{\delta} \quad (451)$$

where \bar{z} , \bar{a} , $\bar{\gamma}$ and $\bar{\delta}$ are given by equations (40), (42), (293) and (294) respectively. The fission instability boundary is given by

$$\bar{z}^2/\bar{a} = \kappa\bar{\gamma}/\bar{\delta} \quad (452)$$

or equivalently the two scalar fission stability boundary conditions are

$$z^2/a = \kappa\gamma/\delta \quad (453)$$

$$\theta_a = 2\theta_z - \theta_\gamma + \theta_\delta \quad (454)$$

Therefore in an external field the internal phase angles of the atomic number, atomic mass number, surface energy coefficient and the Coulomb energy coefficient enter into the fission instability condition. Equations (453) and (454) will now be solved to determine the critical value of θ_z and the critical value of the electromagnetic field strength that are required to catalyze the fission of subactinide nuclei.

B. Critical Value of the Internal Phase Angle of the Atomic Number that is Required for the Low Energy Fission of Subactinide Nuclei in an External Field.

Combining equation (67) with equations (453) and (454) gives the fission instability boundary for nuclei located in an external field as

$$\begin{aligned} z^2/A &= \kappa\gamma/\delta \cos \theta_a \cos^{-2}\theta_z \\ &= \kappa\gamma/\delta \cos(2\theta_z + \theta_\delta - \theta_\gamma) \cos^{-2}\theta_z \end{aligned} \quad (455)$$

Equation (455) must be solved for θ_z . This can be done by noting that simple trigonometry gives

$$\cos(2\theta_z + \theta_\delta - \theta_\gamma) \cos^{-2}\theta_z = (1 - \rho^2) \cos(\theta_\gamma - \theta_\delta) + 2\rho \sin(\theta_\gamma - \theta_\delta) \quad (456)$$

where

$$\rho = \tan \theta_z \quad (457)$$

Equation (455) can then be written as a quadratic equation

$$ap^2 + bp + c = 0 \quad (458)$$

where

$$a = \cos(\theta_Y - \theta_\delta) \quad (459)$$

$$b = 2 \sin(\theta_\delta - \theta_Y) = -2 \sin(\theta_Y - \theta_\delta) \quad (460)$$

$$c = \delta/(\kappa\gamma)Z^2/A - \cos(\theta_Y - \theta_\delta) = \gamma - \cos(\theta_Y - \theta_\delta) \quad (461)$$

Then the critical angle for fission θ_Z^F is given by

$$\begin{aligned} \tan \theta_Z^F &= \tan(\theta_Y - \theta_\delta) \pm \sec(\theta_Y - \theta_\delta) [1 - \delta/(\kappa\gamma)Z^2/A \cos(\theta_Y - \theta_\delta)]^{1/2} \quad (462) \\ &= \tan(\theta_Y - \theta_\delta) \pm \sec(\theta_Y - \theta_\delta) [1 - \chi \cos(\theta_Y - \theta_\delta)]^{1/2} \end{aligned}$$

and from equation (454) the corresponding critical angle for fission θ_a^F is given by

$$\theta_a^F = 2\theta_Z^F + \theta_\delta - \theta_Y \quad (463)$$

The angles θ_Z^F and θ_a^F are the critical values of the phase angles θ_Z and θ_a respectively that are required to bring a nucleus (Z,A) into a state of incipient fission.

Equation (462) is the equation for the instability boundary for the fission of a nucleus in the presence of an external field, and is valid for

$$0 < Z^2/A < \kappa\gamma/\delta \sec(\theta_Y - \theta_\delta) \quad (464A)$$

or in terms of the fissility parameter

$$0 < \chi < \sec(\theta_Y - \theta_\delta) \quad (464B)$$

The values of θ_Z^F corresponding to $\chi = 0$ for the positive and negative modes are given by

$$\tan \theta_{zo}^{F\pm} = \tan(\theta_Y - \theta_\delta) \pm \sec(\theta_Y - \theta_\delta) \quad (465)$$

Using simple trigonometric identities gives

$$\theta_{zo}^{F\pm} = \pm \pi/4 + 1/2(\theta_Y - \theta_\delta) \quad (466)$$

The common value of θ_Z^F and θ_a^F for both positive and negative modes corresponding to $\chi = \sec(\theta_Y - \theta_\delta)$ is given by

$$\theta_z^{Fc} = \theta_a^{Fc} = \theta_Y - \theta_\delta \quad (467)$$

The ranges of variation of θ_Z^F and θ_a^F for the positive and negative angle modes subject to $\chi \leq \sec(\theta_Y - \theta_\delta)$ are given by

$$\theta_z^{Fc} \leq \theta_z^F \leq \theta_{zo}^{F+} \quad \theta_{zo}^{F-} \leq \theta_z^F \leq \theta_z^{Fc} \quad (468)$$

$$\theta_a^{Fc} \leq \theta_a^F \leq \pi/2 \quad -\pi/2 \leq \theta_a^F \leq \theta_a^{Fc} \quad (469)$$

The condition for spontaneous or thermal neutron induced fission in an external field is obtained from equation (462) to be for subactinide nuclei

$$F \leq \chi \leq \sec(\theta_\gamma - \theta_\delta) \quad (470)$$

where χ = fissility parameter, and where

$$F = (1 - \lambda^2) \sec(\theta_\gamma - \theta_\delta) \quad (471)$$

$$\lambda = [\tan \theta_z - \tan(\theta_\gamma - \theta_\delta)] \cos(\theta_\gamma - \theta_\delta) \quad (472)$$

If the external field is shut off all the internal phase angles have zero values and $\chi = 0$ and $F = 1$ so that equation (470) reduces to the result $\chi = 1$ for subactinide nuclei.

As a first approximation the condition $\theta_\gamma = \theta_\delta$ can be taken in equation (463) and the phase angle condition for spontaneous or thermal neutron induced fission in an external field is

$$\theta_a^F = 2\theta_z^F = 2\theta_n^F \quad (473)$$

Combining equations (455) and (473) gives the approximate fission instability boundary for nuclei in an external field as

$$Z^2/A = (\kappa\gamma/\delta)(1 - \tan^2 \theta_z^F) \quad (474)$$

or equivalently as

$$Z^2/A = (\kappa\gamma/\delta)[1 - \tan^2(\theta_a^F/2)] \quad (475)$$

The condition for fission is therefore

$$(Z^2/A)(\kappa\gamma/\delta)^{-1} \geq 1 - \tan^2 \theta_z \quad (476A)$$

or equivalently

$$1 - \tan^2 \theta_z \leq \chi \leq 1 \quad \theta_z \geq \theta_z^F \quad (476B)$$

which follows directly from equations (470) through (472) when $\theta_\gamma = \theta_\delta$. In the presence of an electromagnetic field the fission condition is given by equation (476B), while for a zero external field equation (446) gives the fission condition. From equations (474) through (476) it is clear that in the presence of an external field nuclei lighter than the actinides can fission spontaneously or by thermal neutron absorption unless this is prevented by nuclear shell effects as will be the case for some subactinide nuclei. In general the application of an

electromagnetic or gravitational field tends to lead to more fission instability. The conclusion of this paragraph that subactinide nuclei in the presence of an external field can be fissioned by thermal neutrons depends on the validity of the following conditions

$$\theta_{\gamma} - \theta_{\delta} \ll \theta_z \quad (477)$$

$$\lambda \sim \tan \theta_z \quad (478)$$

$$F \sim 1 - \tan^2 \theta_z \quad (479)$$

where the functions F and λ are given by equations (471) and (472) respectively. Equations (477) through (479) are generally true if $\theta_{\gamma} \sim \theta_{\delta}$, and therefore $F < 1$ in equation (470).

From equation (474) it follows that the internal phase angle of the atomic number that is required for the fission instability of subactinide nuclei is given approximately for the positive angle mode by

$$\begin{aligned} \tan \theta_z^F &= [1 - \delta/(\kappa\gamma)Z^2/A]^{1/2} \\ &= (1 - \chi)^{1/2} \end{aligned} \quad (480)$$

Equation (480) can also be obtained directly from the exact equation (462) by making the approximation $\theta_{\gamma} = \theta_{\delta}$. Equation (480) can also be written as

$$\begin{aligned} \cos \theta_z^F &= [2 - \delta/(\kappa\gamma)Z^2/A]^{-1/2} \\ &= (2 - \chi)^{-1/2} \end{aligned} \quad (481)$$

$$\begin{aligned} \sin \theta_z^F &= \{[1 - \delta/(\kappa\gamma)Z^2/A]/[2 - \delta/(\kappa\gamma)Z^2/A]\}^{1/2} \\ &= [(1 - \chi)/(2 - \chi)]^{1/2} \end{aligned} \quad (482)$$

Equivalently equations (480) and (473) give approximately

$$\begin{aligned} \theta_z^F &= \tan^{-1}[1 - \delta/(\kappa\gamma)Z^2/A]^{1/2} \\ &= \tan^{-1}(1 - \chi)^{1/2} \end{aligned} \quad (483)$$

$$\begin{aligned} \theta_a^F &= 2 \tan^{-1}[1 - \delta/(\kappa\gamma)Z^2/A]^{1/2} \\ &= 2 \tan^{-1}(1 - \chi)^{1/2} \end{aligned} \quad (484)$$

Equations (480) through (482) are valid for $\theta_{\gamma} = \theta_{\delta}$ and the following range of the fissility parameter

$$0 \leq Z^2/A \leq \kappa\gamma/\delta \quad (485)$$

or

$$0 \leq \chi \leq 1 \quad (486)$$

for which the range of values of θ_z^F and θ_a^F are

$$0 \leq \theta_z^F \leq \pi/4 \quad (487)$$

$$0 \leq \theta_a^F \leq \pi/2 \quad (488)$$

which correspond to the exact relations given in equations (468) and (469) respectively. For small values of χ the following approximations are valid

$$\tan \theta_z^F \sim 1 - \chi/2 \quad (489)$$

$$\sin \theta_z^F \sim 2^{-1/2}(1 - \chi/4) \quad (490)$$

$$\cos \theta_z^F \sim 2^{-1/2}(1 + \chi/4) \quad (491)$$

The internal phase angle of the atomic number at incipient fission θ_z^F as given by the approximation in equation (483) is presented in Figure 1.

C. Determination of the Electromagnetic Field Strength Required to Catalyze Fission in Subactinide Nuclei by Thermal Neutrons.

The value of the magnetic field required to bring a lighter than actinide nucleus to the point of fission instability is obtained from equations (209) and (210) as

$$H^F = \theta_z^F / C_{\theta_z}^H = K_{\theta_z}^H \theta_z^F \quad (492)$$

where θ_z^F is given for the general case by equation (462). Therefore in general

$$H^F = H^F(K_{\theta_z}^H, \theta_\gamma - \theta_\delta, \chi) \quad (493)$$

where the fissility parameter χ is given by equation (445). If θ_z^F is given by the approximate equation (483) the critical value of the magnetic field required to catalyze fission in a subactinide nucleus using thermal neutrons is given by

$$\begin{aligned} H^F &= K_{\theta_z}^H \tan^{-1}(1 - \chi)^{1/2} \\ &= K_{\theta_z}^H \tan^{-1}[1 - \delta/(\kappa_\gamma)Z^2/A]^{1/2} \end{aligned} \quad (494)$$

so that within this approximation

$$H^F = H^F(K_{\theta_z}^H, \chi) \quad (495)$$

Equation (494) is valid for $\chi \leq 1$.

The condition $\chi < 1$ generally occurs in subactinide nuclei or in the actinides with large neutron excess, while the condition $\chi > 1$ generally occurs in the actinide and transactinide nuclei. Equation (494) gives the magnetic field strength required to catalyze spontaneous or thermal neutron induced fission in nuclei for which $\chi < 1$ except in those nuclei where nuclear shell effects add increased stability and do not allow fission to occur. For the case when $\chi > 1$ nuclei will generally fission spontaneously or by thermal neutron induced fission without the presence of an electromagnetic field except for those nuclei where shell effects give increased stability against fission as for example in the case of ^{238}U . If the value of K_{0z}^H is very large the magnetic field strength required to catalyze clean fission in the subactinide nuclei will be too large for laboratory demonstration and practical nuclear reactor design. In Section 7 it will be shown that the static magnetic field required to catalyze clean fission in the subactinide elements is in the teratesla range and is too large for practical purposes. However, in Section 7 it is also shown that the application of a γ ray electromagnetic field with a relatively low magnetic field vector strength can catalyze clean fission in the subactinide elements using thermal neutrons but only when nuclear shell stability does not prohibit the occurrence of the phenomenon.

6. FINAL STATE ENERGY CONDITIONS FOR THE FISSION OF NUCLEI IN AN EXTERNAL FIELD. This section considers a comparison between the initial and final energy states of an atomic nucleus that has fissioned in the presence of an external electromagnetic or gravitational field. A fission reaction in which a nucleus (Z, A) has split into two nuclei (Z_1, A_1) and (Z_2, A_2) is written in the form⁶⁹⁻⁸⁰

$$(Z, A) \rightarrow (Z_1, A_1) + (Z_2, A_2) \quad (496)$$

Then the nucleus (Z_2, A_2) is assumed to eject a neutron

$$(Z_2, A_2) \rightarrow (Z_2, A_2 - 1) + (0, 1) \quad (497)$$

where in this notation $(0, 1)$ is a single neutron. In this way the general process of nuclear fission can be represented by a nuclear transformation of the general form given in equation (496). In an external field the nuclei represented in equation (496) are also associated with complex atomic numbers, neutron numbers and atomic mass numbers that in analogy to equations (40) through (42) are represented for subactinide nuclei by

$$\bar{z} = z \exp(j\theta_z) = Z \cos \theta_z \exp(j\theta_z) \quad (498)$$

$$\bar{n} = n \exp(j\theta_n) = N \cos \theta_n \exp(j\theta_n) \quad (499)$$

$$\bar{a} = a \exp(j\theta_a) = A \cos \theta_a \exp(j\theta_a) \quad (500)$$

$$\bar{z}_1 = z_1 \exp(j\theta_{z1}) = Z_1 \cos \theta_{z1} \exp(j\theta_{z1}) \quad (501)$$

$$\bar{n}_1 = n_1 \exp(j\theta_{n1}) = N_1 \cos \theta_{n1} \exp(j\theta_{n1}) \quad (502)$$

$$\bar{a}_1 = a_1 \exp(j\theta_{a1}) = A_1 \cos \theta_{a1} \exp(j\theta_{a1}) \quad (503)$$

$$\bar{z}_2 = z_2 \exp(j\theta_{z2}) = Z_2 \cos \theta_{z2} \exp(j\theta_{z2}) \quad (504)$$

$$\bar{n}_2 = n_2 \exp(j\theta_{n2}) = N_2 \cos \theta_{n2} \exp(j\theta_{n2}) \quad (505)$$

$$\bar{a}_2 = a_2 \exp(j\theta_{a2}) = A_2 \cos \theta_{a2} \exp(j\theta_{a2}) \quad (506)$$

The determination of the energy released during the fission reaction given in equation (496) requires that all of the nine internal phase angles that appear in equations (498) through (506) be determined, and the procedure for doing this will now be given.

A. Determination of the Internal Phase Angles of the Atomic Number, Neutron Number and Atomic Mass Number for the Initial and Final States of a Fission Reaction for Subactinide Nuclei.

The nuclei involved in the fission reaction given by equation (496) are subject to the following scalar baryon number conservation equations

$$A = Z + N \quad A_1 = Z_1 + N_1 \quad A_2 = Z_2 + N_2 \quad (507)$$

$$A = A_1 + A_2 \quad Z = Z_1 + Z_2 \quad N = N_1 + N_2 \quad (508)$$

In an external field the nuclei represented by equation (496) are subject to the following complex atomic number, neutron number and atomic mass number conservation equations similar to equation (82)

$$\bar{a} + W = \bar{z} + \bar{n} \quad \bar{a}_1 + W_1 = \bar{z}_1 + \bar{n}_1 \quad \bar{a}_2 + W_2 = \bar{z}_2 + \bar{n}_2 \quad (509)$$

$$\bar{a} + W_a = \bar{a}_1 + \bar{a}_2 \quad \bar{z} + W_z = \bar{z}_1 + \bar{z}_2 \quad \bar{n} + W_n = \bar{n}_1 + \bar{n}_2 \quad (510)$$

Equations (509) and (510) show that all of the W's are not independent, and in fact they are subject to the following equation

$$W - W_1 - W_2 = W_a - W_z - W_n \quad (511)$$

Equations (509) and (510) can be combined with equations (498) through (506) to yield the following twelve equations for subactinide nuclei

$$A \cos^2 \theta_a + W = Z \cos^2 \theta_z + N \cos^2 \theta_n \quad (512)$$

$$A \cos \theta_a \sin \theta_a = Z \cos \theta_z \sin \theta_z + N \cos \theta_n \sin \theta_n \quad (513)$$

$$A_1 \cos^2 \theta_{a1} + W_1 = Z_1 \cos^2 \theta_{z1} + N_1 \cos^2 \theta_{n1} \quad (514)$$

$$A_1 \cos \theta_{a1} \sin \theta_{a1} = Z_1 \cos \theta_{z1} \sin \theta_{z1} + N_1 \cos \theta_{n1} \sin \theta_{n1} \quad (515)$$

$$A_2 \cos^2 \theta_{a2} + W_2 = Z_2 \cos^2 \theta_{z2} + N_2 \cos^2 \theta_{n2} \quad (516)$$

$$A_2 \cos \theta_{a2} \sin \theta_{a2} = Z_2 \cos \theta_{z2} \sin \theta_{z2} + N_2 \cos \theta_{n2} \sin \theta_{n2} \quad (517)$$

$$A \cos^2 \theta_a + W_a = A_1 \cos^2 \theta_{a1} + A_2 \cos^2 \theta_{a2} \quad (518)$$

$$A \cos \theta_a \sin \theta_a = A_1 \cos \theta_{a1} \sin \theta_{a1} + A_2 \cos \theta_{a2} \sin \theta_{a2} \quad (519)$$

$$Z \cos^2 \theta_z + W_z = Z_1 \cos^2 \theta_{z1} + Z_2 \cos^2 \theta_{z2} \quad (520)$$

$$Z \cos \theta_z \sin \theta_z = Z_1 \cos \theta_{z1} \sin \theta_{z1} + Z_2 \cos \theta_{z2} \sin \theta_{z2} \quad (521)$$

$$N \cos^2 \theta_n + W_n = N_1 \cos^2 \theta_{n1} + N_2 \cos^2 \theta_{n2} \quad (522)$$

$$N \cos \theta_n \sin \theta_n = N_1 \cos \theta_{n1} \sin \theta_{n1} + N_2 \cos \theta_{n2} \sin \theta_{n2} \quad (523)$$

where Z, N, A ; Z_1, N_1, A_1 and Z_2, N_2, A_2 are known quantities.

There are fifteen unknown quantities in the problem of the fission of an atomic nucleus in the presence of an electromagnetic field:

$$W, \theta_z, \theta_n, \theta_a \quad (524)$$

$$W_1, \theta_{z1}, \theta_{n1}, \theta_{a1} \quad (525)$$

$$W_2, \theta_{z2}, \theta_{n2}, \theta_{a2} \quad (526)$$

$$W_z, W_n, W_a \quad (527)$$

There are fifteen equations to determine these quantities and they are: the twelve equations (512) through (523), the two fission instability equations (453) and (454) which determine θ_z and θ_a in the forms of equations (462) and (463), and finally equation (511) which relates the various W -functions. The values of the W 's are similar in form to the functions required for the addition law of complex magnetic quantum numbers as given in equation (37), and are written as follows for subactinide nuclei

$$W = -A/2[1 + (1 - 4f_W^2)^{1/2}] + Z \cos^2 \theta_z + N \cos^2 \theta_n \quad (528)$$

$$W_1 = -A_1/2[1 + (1 - 4f_{W1}^2)^{1/2}] + Z_1 \cos^2 \theta_{z1} + N_1 \cos^2 \theta_{n1} \quad (529)$$

$$W_2 = -A_2/2[1 + (1 - 4f_{W2}^2)^{1/2}] + Z_2 \cos^2 \theta_{z2} + N_2 \cos^2 \theta_{n2} \quad (530)$$

$$W_z = -Z/2[1 + (1 - 4f_{Wz}^2)^{1/2}] + Z_1 \cos^2 \theta_{z1} + Z_2 \cos^2 \theta_{z2} \quad (531)$$

$$W_n = -N/2[1 + (1 - 4f_{Wn}^2)^{1/2}] + N_1 \cos^2 \theta_{n1} + N_2 \cos^2 \theta_{n2} \quad (532)$$

$$W_a = -A/2[1 + (1 - 4f_{Wa}^2)^{1/2}] + A_1 \cos^2 \theta_{a1} + A_2 \cos^2 \theta_{a2} \quad (533)$$

where

$$f_W = A^{-1}(Z \sin \theta_z \cos \theta_z + N \sin \theta_n \cos \theta_n) \quad (534)$$

$$f_{W1} = A_1^{-1}(Z_1 \sin \theta_{z1} \cos \theta_{z1} + N_1 \sin \theta_{n1} \cos \theta_{n1}) \quad (535)$$

$$f_{W2} = A_2^{-1}(Z_2 \sin \theta_{z2} \cos \theta_{z2} + N_2 \sin \theta_{n2} \cos \theta_{n2}) \quad (536)$$

$$f_{Wz} = Z^{-1}(Z_1 \sin \theta_{z1} \cos \theta_{z1} + Z_2 \sin \theta_{z2} \cos \theta_{z2}) \quad (537)$$

$$f_{Wn} = N^{-1}(N_1 \sin \theta_{n1} \cos \theta_{n1} + N_2 \sin \theta_{n2} \cos \theta_{n2}) \quad (538)$$

$$f_{Wa} = A^{-1}(A_1 \sin \theta_{a1} \cos \theta_{a1} + A_2 \sin \theta_{a2} \cos \theta_{a2}) \quad (539)$$

and where

$$\cos^2 \theta_a = 1/2[1 + (1 - 4f_W^2)^{1/2}] \quad (540)$$

$$\cos^2 \theta_{a1} = 1/2[1 + (1 - 4f_{W1}^2)^{1/2}] \quad (541)$$

$$\cos^2 \theta_{a2} = 1/2[1 + (1 - 4f_{W2}^2)^{1/2}] \quad (542)$$

$$\cos^2 \theta_z = 1/2[1 + (1 - 4f_{Wz}^2)^{1/2}] \quad (543)$$

$$\cos^2 \theta_n = 1/2[1 + (1 - 4f_{Wn}^2)^{1/2}] \quad (544)$$

$$\cos^2 \theta_a = 1/2[1 + (1 - 4f_{Wa}^2)^{1/2}] \quad (545)$$

Therefore equations (528) through (533) can be written as

$$\begin{aligned} W &= -A \cos^2 \theta_a + Z \cos^2 \theta_z + N \cos^2 \theta_n \\ &= Z(\cos^2 \theta_z - \cos^2 \theta_a) + N(\cos^2 \theta_n - \cos^2 \theta_a) \end{aligned} \quad (546)$$

$$\begin{aligned}
W_1 &= -A_1 \cos^2 \theta_{a1} + Z_1 \cos^2 \theta_{z1} + N_1 \cos^2 \theta_{n1} \\
&= Z_1 (\cos^2 \theta_{z1} - \cos^2 \theta_{a1}) + N_1 (\cos^2 \theta_{n1} - \cos^2 \theta_{a1})
\end{aligned} \tag{547}$$

$$\begin{aligned}
W_2 &= -A_2 \cos^2 \theta_{a2} + Z_2 \cos^2 \theta_{z2} + N_2 \cos^2 \theta_{n2} \\
&= Z_2 (\cos^2 \theta_{z2} - \cos^2 \theta_{a2}) + N_2 (\cos^2 \theta_{n2} - \cos^2 \theta_{a2})
\end{aligned} \tag{548}$$

$$\begin{aligned}
W_z &= -Z \cos^2 \theta_z + Z_1 \cos^2 \theta_{z1} + Z_2 \cos^2 \theta_{z2} \\
&= Z_1 (\cos^2 \theta_{z1} - \cos^2 \theta_z) + Z_2 (\cos^2 \theta_{z2} - \cos^2 \theta_z)
\end{aligned} \tag{549}$$

$$\begin{aligned}
W_n &= -N \cos^2 \theta_n + N_1 \cos^2 \theta_{n1} + N_2 \cos^2 \theta_{n2} \\
&= N_1 (\cos^2 \theta_{n1} - \cos^2 \theta_n) + N_2 (\cos^2 \theta_{n2} - \cos^2 \theta_n)
\end{aligned} \tag{550}$$

$$\begin{aligned}
W_a &= -A \cos^2 \theta_a + A_1 \cos^2 \theta_{a1} + A_2 \cos^2 \theta_{a2} \\
&= A_1 (\cos^2 \theta_{a1} - \cos^2 \theta_a) + A_2 (\cos^2 \theta_{a2} - \cos^2 \theta_a)
\end{aligned} \tag{551}$$

A comparison of equations (540) and (545) shows that

$$f_W = f_{Wa} \tag{552}$$

and equations (546) and (551) give

$$W - W_a = Z \cos^2 \theta_z + N \cos^2 \theta_n - A_1 \cos^2 \theta_{a1} - A_2 \cos^2 \theta_{a2} \tag{553}$$

Equation (473) shows that a nucleus in a state of incipient fission has $\theta_a \sim 2\theta_z$ and therefore equation (546) gives $W > 0$. In general for fission in an external field where the internal phase angles are relatively large

$$W > 0 \qquad W_1 > 0 \qquad W_2 > 0 \tag{554}$$

$$W_z < 0 \qquad W_n < 0 \qquad W_a < 0 \tag{555}$$

When the external electromagnetic or gravitational field is shut off all of the W 's have zero values. However, in reality external fields are always present.

B. Energy Released from Nuclear Fission in an External Field.

The Q value of a nuclear reaction is a measure of the energy released in a nuclear fission process.⁶¹ In this paper a complex number generalization of the standard definition of the Q value is given by

$$\bar{Q}/c^2 = \bar{M}(A, Z) - \bar{M}(A_1, Z_1) - \bar{M}(A_2, Z_2) \quad (556)$$

where as in equation (361)

$$\bar{M}(A, Z) = \bar{n}m_n + \bar{z}m_H - \bar{B}(A, Z)/c^2 \quad (557)$$

$$\bar{M}(A_1, Z_1) = \bar{n}_1m_n + \bar{z}_1m_H - \bar{B}(A_1, Z_1)/c^2 \quad (558)$$

$$\bar{M}(A_2, Z_2) = \bar{n}_2m_n + \bar{z}_2m_H - \bar{B}(A_2, Z_2)/c^2 \quad (559)$$

Then the \bar{Q} value can be written as

$$\begin{aligned} \bar{Q} = & [(\bar{n} - \bar{n}_1 - \bar{n}_2)m_n + (\bar{z} - \bar{z}_1 - \bar{z}_2)m_H]c^2 \\ & + \bar{B}(A_1, Z_1) + \bar{B}(A_2, Z_2) - \bar{B}(A, Z) \end{aligned} \quad (560)$$

Using equation (510) allows equation (560) to be written as

$$\bar{Q} = Q_1 + \bar{Q}_2 \quad (561)$$

where

$$Q_1 = - (W_n m_n + W_z m_H)c^2 \quad (562)$$

$$\bar{Q}_2 = \bar{B}(A_1, Z_1) + \bar{B}(A_2, Z_2) - \bar{B}(A, Z) \quad (563)$$

where W_z and W_n are given by equations (549) and (550) respectively.

Because $W_n < 0$ and $W_z < 0$ it follows that for subactinide nuclei

$$Q_1 > 0 \quad (564)$$

The value of Q_1 arises from the rest mass terms in equations (557) through (562). The actual rest mass is unchanged in a nuclear fission process because

$$Nm_n + Zm_H - (N_1m_n + Z_1m_H) - (N_2m_n + Z_2m_H) = 0 \quad (565)$$

which is always true because of the absolute validity of baryon number conservation which for the present case is written as

$$Z = Z_1 + Z_2 \quad N = N_1 + N_2 \quad (566)$$

A finite value of Q_1 results from the special form of the conservation law of complex baryon numbers, which for the complex atomic number, neutron number and atomic mass number are given in equations (509) and (510). The nonzero value of Q_1 does not represent a violation of the law of baryon number conservation but instead is a manifestation of the broken symmetry of the atomic number and the neutron number in an external field. Note that the expression for Q_1 can be rewritten using equations (549), (550) and (562) as

$$Q_1/c^2 = -m_n[N_1(\cos^2\theta_{n1} - \cos^2\theta_n) + N_2(\cos^2\theta_{n2} - \cos^2\theta_n)] \\ - m_H[Z_1(\cos^2\theta_{z1} - \cos^2\theta_z) + Z_2(\cos^2\theta_{z2} - \cos^2\theta_z)] \quad (567)$$

In general Q_1 can be taken to be a small number and can be neglected compared to the value of \bar{Q}_2 . For zero value of the applied external field $Q_1 = 0$ because all internal phase angles have zero values, and therefore $W_n = 0$ and $W_z = 0$.

The value of \bar{Q}_2 can be calculated by combining equations (292) and (563). This is easily done for symmetric fission and under the approximation

$$\theta_{z1} \sim \theta_{z2} \sim \theta_z \quad \theta_{n1} \sim \theta_{n2} \sim \theta_n \quad (568)$$

For symmetric fission equation (563) becomes

$$\bar{Q}_2 = 2\bar{B}(A/2, Z/2) - \bar{B}(A, Z) \quad (569)$$

Under these assumptions the value of \bar{Q}_2 is given by the following complex number generalization of the standard scalar result⁶¹

$$\bar{Q}_2 = (1 - 2^{1/3})\bar{\gamma}\bar{a}^{2/3} + (1 - 2^{-2/3})\bar{\delta}\bar{z}^2/\bar{a}^{1/3} \\ = -0.26\bar{\gamma}\bar{a}^{2/3} + 0.37\bar{\delta}\bar{z}^2/\bar{a}^{1/3} \quad (570)$$

The simple form in equation (570) results from the approximation given in equation (568). The value of \bar{Q} is then written as

$$\bar{Q} = - (W_n m_n + W_z m_H) c^2 - 0.26\bar{\gamma}\bar{a}^{2/3} + 0.37\bar{\delta}\bar{z}^2/\bar{a}^{1/3} \quad (571)$$

The measured value of \bar{Q} is given by the real part of equation (571)

$$Q_m = - (W_n m_n + W_z m_H) c^2 - 0.26\gamma A^{2/3} \cos^{2/3}\theta_a \cos(\theta_\gamma + 2/3\theta_a) \\ + 0.37\delta Z^2 A^{-1/3} \cos^2\theta_z \cos^{-1/3}\theta_a \cos(\theta_\delta + 2\theta_z - 1/3\theta_a) \quad (572)$$

Equation (572) can be compared to the conventionally calculated value of Q which is given by⁶¹

$$Q_c = -0.26\gamma A^{2/3} + 0.37\delta Z^2 A^{-1/3} \quad (573)$$

For the case of zero external field equation (572) reduces to equation (573).

A condition that determines the possibility of the final fission state to occur can be obtained from the Q value for the nuclear fission process.⁶¹ The complex number generalization of this condition is

$$\bar{Q} \geq \bar{E}_c^* \quad (574)$$

where \bar{E}_c^* = complex number Coulomb potential energy of two spherical nuclei (Z/2, A/2) in geometrical contact. This Coulomb energy can be written as a simple complex number generalization of the standard scalar result⁶¹

$$\begin{aligned} \bar{E}_c^* &= 1/2 e^2 (\bar{z}/2)^2 / [\bar{b}(\bar{a}/2)^{1/3}] \\ &= 2^{1/3} (1/8) (5/3) \bar{\delta} \bar{z}^2 / \bar{a}^{1/3} = 0.262 \bar{\delta} \bar{z}^2 / \bar{a}^{1/3} \end{aligned} \quad (575)$$

where as before in equation (329)

$$\bar{\delta} = 3/5 e^2 / \bar{b} = 0.863 / \bar{b} = (0.863 / 1.523) \bar{k}_c \quad \text{MeV} \quad (576)$$

where \bar{b} = complex number radius parameter given by equation (283). For $k_c = 1.35 \text{ fm}^{-1}$ as in equation (332) it follows that

$$\delta = 0.765 \text{ MeV} = 765 \text{ keV} \quad (577)$$

and θ_δ is given by equation (331).

Combining equations (571), (574) and (575) gives the final state fission energy condition as

$$(W_{nn} + W_{zH})c^2 + 0.26\gamma\bar{a}^{2/3} = 0.11\bar{\delta}\bar{z}^2/\bar{a}^{1/3} \quad (578)$$

This equation can be used instead of the incipient fission condition given in equation (452) to determine θ_z and θ_a . However because of the presence of the functions W_n and W_z the full set of thirteen equations (511) through (523) must be solved in conjunction with the two components of equation (578) which are

$$(W_{nn} + W_{zH})c^2 + 0.26\gamma\bar{a}^{2/3} \cos^{2/3}\theta_a \cos(\theta_\gamma + 2/3\theta_a) \quad (579)$$

$$= 0.11\bar{\delta}\bar{z}^2\bar{a}^{-1/3} \cos^2\theta_z \cos^{-1/3}\theta_a \cos(\theta_\delta + 2\theta_z - 1/3\theta_a)$$

$$0.26\gamma\bar{a}^{2/3} \cos^{2/3}\theta_a \sin(\theta_\gamma + 2/3\theta_a) \quad (580)$$

$$= 0.11\bar{\delta}\bar{z}^2\bar{a}^{-1/3} \cos^2\theta_z \cos^{-1/3}\theta_a \sin(\theta_\delta + 2\theta_z - 1/3\theta_a)$$

If W_n and W_z are neglected in equation (578) then the final state fission condition can be written as

$$\bar{z}^2/\bar{a} = \kappa' \bar{\gamma}/\bar{\delta} \quad \kappa' = 2.36 \quad (581)$$

Equation (581) is the same form as the incipient fission condition given in equation (451) and the same form of solution for θ_Z^F and θ_a^F that appears in equations (462) and (480) can now be used to determine these phase angles for the final state fission condition. Then the remaining thirteen equations (511) through (523) can be used to calculate the remaining thirteen functions listed in equations (524) through (527).

7. SUSTAINED γ RAY CATALYZED THERMAL NEUTRON INDUCED CLEAN FISSION NUCLEAR REACTIONS. This section presents numerical calculations of the internal phase angle of the atomic number θ_Z^F that is required for the external field catalysis of spontaneous or thermal neutron induced fission nuclear reactions in nuclei lighter than the actinides. The corresponding strengths of the static magnetic field and electromagnetic wave field required to catalyze spontaneous or thermal neutron induced fission in subactinide nuclei are also obtained. Under ordinary conditions these relatively light nuclei are not fissile for incident thermal neutrons. In fact, under ordinary circumstances only some of the actinides sustain fission by thermal neutrons, for example ^{235}U and ^{239}Pu are fissile for incident thermal neutrons but ^{238}U and ^{232}Th are not. Higher incident neutron energies will induce fission in all of the actinides, and in fact light elements will undergo neutron induced fission for sufficiently high kinetic energy of the incident neutrons - the nuclei are simply blown apart. For nuclear reactors requiring sustained fission reactions, however, thermal neutron induced fission reactions are required so that at present only actinide elements such as ^{235}U and ^{239}Pu can be used. The fission products of these heavy nuclei are dangerous radionuclides.

The fission product nuclei from conventional actinide element nuclear reactors are radioactive because they have large neutron excesses and are far removed from the valley of beta stability. The primary radiations from conventional nuclear reactor fission products are beta decays, alpha decays and neutron emissions which occur as the nuclei move toward the valley of beta stability. The only way to have clean low energy fission is to have fission product nuclei that are close to the valley of beta stability, and the only practical way to achieve this for nuclear reactors is to have thermal neutron induced fission in nuclei lighter than the actinides. Section 5 of this paper showed that this is possible if the subactinide nuclei are immersed in an electromagnetic field. Many other surprising effects occur when atoms are immersed in an electromagnetic field.⁸¹⁻⁹⁰ This section develops the electromagnetic field criteria that are necessary for catalyzing the clean fission of relatively light subactinide nuclei using thermal neutrons, and in particular it is shown that a γ ray field is required. The fission products of these reactions are either not radioactive or exhibit only low level emissions. Examples of the clean fission reactions for subactinide nuclei are presented.

A. Numerical Values of the Internal Phase Angle θ_Z^F Required for the Low Energy Clean Fission of Subactinide Nuclei.

The general expression for θ_Z^F associated with the fission of a subactinide nucleus (A,Z) requires θ_γ and θ_δ as input parameters as shown in equation (462). These two mass formula parameters can only be determined by fitting the

broken symmetry form of the liquid drop nuclear mass formula given by equations (347) and (369) to measured nuclear mass data. As this was not done, the assumption $\theta_\gamma = \theta_\delta$ is made and only the approximate expression for θ_z^F given by equation (483) is used for numerical analysis. Table 1 gives the values of θ_z^F required for thermal neutron induced fission. The results in Table 1 are calculated for $\kappa = 1.471$, $\gamma = 17.2$ MeV, $\delta = 0.698$ MeV and $\kappa\gamma/\delta = 36.25$ which enters the fissility parameter calculation in equation (445). The average kinetic energy of the thermal neutrons is $\epsilon_k = 0.025$ eV.

The nuclei in Table 1 are arranged by decreasing values of the fissility parameter. Those actinide nuclei for which thermal neutron induced fission occurs without the need of an external electromagnetic field are indicated by zero values of θ_z^F . The fact that there is no sharp boundary between nuclei that can be fissioned with thermal neutrons and those that cannot shows that the Bohr-Wheeler fissility parameter is a collective property of a nucleus and does not completely describe the nuclear fission process because nuclear shell effects play an important role. For a gross description of nuclear fission the fissility parameter is adequate. Table 1 shows that the angle θ_z^F is relatively small for the actinide nuclei that require an external electromagnetic field to undergo fission by thermal neutrons such as ^{234}U , ^{231}Pa , ^{238}U and ^{232}Th , so that any experimental confirmation of electromagnetically catalyzed clean fission would be easiest for these actinide nuclei unless strong nuclear shell effects enter to prevent low energy fission.

B. Static Magnetic Field Required to Catalyze Low Energy Clean Fission in Subactinide Nuclei.

The determination of the static magnetic field necessary to catalyze clean fission in nuclei lighter than the actinides using thermal neutrons requires values of the magnetic stiffness coefficients $K_{\theta z}^B$ and $K_{\theta z}^H$ or equivalently the magnetic compliance coefficients $C_{\theta z}^B$ and $C_{\theta z}^H$ that are defined in equations (185) through (214). The values of these coefficients can be calculated by requiring the equality of the potential energy density of elastic shearing in internal space and the magnetic energy density of the applied field. This equality is written as

$$1/2 G \theta_z^2 = B^2 / (2\mu) = 1/2 \mu H^2 \quad (582)$$

or as in equations (209) through (212)

$$\theta_z = C_{\theta z}^B B \quad B = K_{\theta z}^B \theta_z \quad (583)$$

$$\theta_z = C_{\theta z}^H H \quad H = K_{\theta z}^H \theta_z \quad (584)$$

where from equation (582)

$$C_{\theta z}^B = (\mu G)^{-1/2} \quad C_{\theta z}^H = (\mu/G)^{1/2} \quad (585)$$

$$K_{\theta z}^B = (\mu G)^{1/2} \quad K_{\theta z}^H = (G/\mu)^{1/2} \quad (586)$$

where G and μ = shear modulus and magnetic permeability of an atomic nucleus, and B and H = magnitudes of the magnetic induction vector and magnetic field vector respectively. The values of B and H required to fission a subactinide nucleus are then given by equations (583), (584) and (586) as

$$B^F = K_{\theta z}^B \theta_z^F = (G\mu)^{1/2} \theta_z^F \quad (587)$$

$$H^F = K_{\theta z}^H \theta_z^F = (G/\mu)^{1/2} \theta_z^F \quad (588)$$

so that generally if equation (462) is used for θ_z^F

$$B^F = B^F(\theta_\gamma, \theta_\delta, \chi, G, \mu) \quad (589)$$

or more simply if equation (483) is used to calculate θ_z^F

$$B^F = B^F(\chi, G, \mu) \quad (590)$$

where χ = fissility parameter given by equation (445). The units of the relevant physical quantities used in this analysis are given by

$$[H] = \text{amp/m} = \text{coul}/(\text{m sec})$$

$$[B] = T = \text{Wb/m}^2 = \text{kg}/(\text{sec coul}) = 10^4 \text{ gauss} = \text{N sec}/(\text{m coul})$$

$$[G] = \text{N/m}^2$$

$$[\mu] = \text{kg m/coul}^2 = \text{Henry/m} = \text{Wb}/(\text{amp m}) = \text{N sec}^2/\text{coul}^2$$

$$[C_{\theta z}^H] = \text{rad m sec/coul} = \text{rad m/amp}$$

$$[C_{\theta z}^B] = \text{rad/T} = \text{rad sec coul/kg} = \text{rad m}^2/\text{Wb}$$

$$[K_{\theta z}^H] = \text{coul}/(\text{rad m sec}) = \text{amp}/(\text{rad m})$$

$$[K_{\theta z}^B] = \text{T/rad} = \text{kg}/(\text{rad sec coul}) = \text{Wb}/(\text{rad m}^2) = \text{N sec}/(\text{rad m coul})$$

where T = tesla.

Equation (587) shows that the value of the magnetic induction field B^F required for clean fission depends on the value of the magnetic permeability of nuclear matter. The value of μ refers to neutrons and protons in nuclear matter at nuclear density which corresponds to $k_c \sim 1.35 \text{ fm}^{-1}$. The vacuum value of μ is given by^{91,92}

$$\mu_0 = 4\pi \times 10^{-7} \text{ kg m/coul}^2 \quad (591)$$

The value of μ corresponding to nuclear matter is given by^{91,92}

$$\mu = \mu_0(1 + \chi_m) \quad (592)$$

where the magnetic susceptibility χ_m of nuclear matter is given by

$$\chi_m = n_c \beta = n_c (\beta^p + \beta^n) \quad (593)$$

where n_c = central nucleon number density of an atomic nucleus which is approximately equal to the saturation nucleon number density of infinite nuclear matter and is given by^{60,68}

$$n_c = 2/(3\pi^2)k_c^3 \sim 0.166 \text{ fm}^{-3} \quad (594)$$

and β^p = magnetic polarizability of the proton and β^n = magnetic polarizability of the neutron which are given by^{93,94}

$$\beta^p \sim \beta^n = 3.34 \times 10^{-4} \text{ fm}^3 \quad (595)$$

Equations (593) through (595) give the magnetic susceptibility of nuclear matter at nuclear density a value of

$$\chi_m = 1.11 \times 10^{-4} \quad (596)$$

so that $\chi_m \ll 1$ and equation (592) gives $\mu \sim \mu_0$, and therefore the vacuum value of the magnetic permeability is adequate for calculating the magnetic properties of nuclear matter at nuclear density.

Equation (587) shows that the value of the magnetic induction field B^F required for the electromagnetic catalysis of thermal neutron induced clean fission of elements lighter than the actinides depends on the shear modulus of nuclear matter. The shear modulus of nuclear matter is easily obtained from the value of the bulk modulus (incompressibility) and Poisson's ratio of nuclear matter as follows⁹⁵

$$G = 3(1 - 2\nu)K[2(1 + \nu)]^{-1} \quad (597)$$

where G , ν and K = shear modulus, Poisson's ratio and bulk modulus of nuclear matter. The values of Poisson's ratio usually have a range of values $0 < \nu < 1/2$, so that a value $\nu = 1/4$ is adopted in this paper. The bulk modulus is given by the following standard formula⁹⁵

$$\begin{aligned} K &= ndP/dn \\ &= n(2nd\epsilon/dn + k) \end{aligned} \quad (598)$$

where n = nucleon number density, P = pressure, ϵ = average energy per nucleon and k = incompressibility parameter given by^{68,96}

$$k = n^2 d^2 \epsilon / dn^2 \quad (599)$$

At the saturation density which occurs at the minimum average energy per nucleon $d\epsilon/dn = 0$, and equation (598) gives the bulk modulus as

$$K = nk = n^3 d^2 \epsilon / dn^2 \quad (600)$$

For nuclear matter with $Z/N = 1$ the following are commonly used values of the saturation nucleon number density and the incompressibility factor^{68,96}

$$n = 0.166 \text{ nucleons/fm}^3 \quad (601)$$

$$k = 250 \text{ MeV/nucleon} \quad (602)$$

Equation (602) gives a relatively high value for the incompressibility factor so that the author will not be accused of ignoring the hard facts.

Combining equations (597) and (600) through (602) gives the bulk modulus and shear modulus of $Z = N$ nuclear matter at nuclear density as

$$K = 41.5 \text{ MeV/fm}^3 = 6.64 \times 10^{33} \text{ N/m}^2 \quad (603)$$

$$G = 24.9 \text{ MeV/fm}^3 = 3.98 \times 10^{33} \text{ N/m}^2 \quad (604)$$

Also, at equilibrium the average binding energy for infinite nuclear matter and for an atomic nucleus with $N = Z$ are respectively

$$\epsilon \sim 15.5 \text{ MeV/nucleon} \quad \text{infinite nuclear matter} \quad (605)$$

$$\epsilon \sim 8.0 \text{ MeV/nucleon} \quad \text{atomic nucleus} \quad (606)$$

while the average Coulomb energy per nucleon is⁶⁸

$$\epsilon_c \sim 0.765 \text{ MeV/nucleon} \quad (607)$$

For purposes of comparison with chemically bound systems it should be pointed out that the shear modulus of steel is⁹⁵

$$G_{\text{steel}} \sim 4 \times 10^9 \text{ N/m}^2 \quad (608)$$

while the binding energy (electronegativity) of a valence electron of a chemical element typically has a value⁹⁷

$$\epsilon_{\text{chem}} \sim 10 \text{ eV/electron} \quad (609)$$

A comparison of equations (605), (606) and (609) shows that the average binding energy per particle in a nuclear system is about 6 orders of magnitude larger than the average binding energy per electron in an atomic system. However, a comparison of equations (604) and (608) shows that the bulk modulus of nuclear matter at nuclear density is about 24 orders of magnitude larger than the bulk modulus of a chemical compound at its equilibrium density. The very large difference in the values of the bulk moduli of nuclear and chemical systems is due to the large value of the nuclear matter density which enters directly into the calculation of the bulk modulus as shown in equation (600). The values of the nuclear binding energy, saturation density and bulk modulus actually depend on the values of the neutron excess of nuclear matter, but this is not considered in this paper.⁶⁸

The magnetic compliance coefficients for nuclear matter in a static mag-

netic field can be calculated from equations (585), (591) and (604), and the results are

$$C_{\theta z}^B = 1.41 \times 10^{-14} \text{ rad/T} = 8.08 \times 10^{-13} \text{ deg/T} \quad (610)$$

$$C_{\theta z}^H = 1.78 \times 10^{-20} \text{ rad m sec/coul} \quad (611)$$

where T = tesla. The magnetic shear stiffness coefficients can be calculated from equations (586), (591) and (604) or more simply as the reciprocals of the magnetic compliance coefficients given in equations (610) and (611) with the result that

$$K_{\theta z}^B = 7.07 \times 10^{13} \text{ T/rad} = 1.23 \times 10^{12} \text{ T/deg} \quad (612)$$

$$K_{\theta z}^H = 5.63 \times 10^{19} \text{ coul/(rad m sec)} \quad (613)$$

The values of the magnetic compliance and magnetic shear stiffness coefficients that have been calculated refer to a static magnetic field because the shear modulus used in these calculations, and given by equation (604), is a measure of the static deformation of nuclear matter in internal space under the application of a static magnetic field as described by equation (582).

The values of the static magnetic induction field required for clean fission can be calculated from equations (587), (612) and the values of θ_z^F given by equation (483). The values of θ_z^F and B^F required for the clean fission of some nuclei of interest appear in Table 1. From Table 1 it is obvious that very large values, ~ 40 teratesla, of the static magnetic induction field B^F are required to catalyze clean fission nuclear reactions in the subactinide nuclei. These values of B^F are much larger than any static magnetic field that has been obtained in the laboratory. Typical static magnetic fields used in particle physics experiments are about 4-6 T.⁹⁸ The highest static magnetic field obtained in the laboratory is about 50 T.^{99,100} Therefore the prediction of static magnetically catalyzed clean fission of subactinide nuclei by thermal neutrons cannot be verified in the laboratory with present day technology. Explosive compression generated magnetic fields have values that exceed several kilotesla but fall far short of the 10^{12} tesla field required for practical clean fission reactions.^{101,104} The huge values of the static magnetic field required for the catalysis of clean fission of nuclei lighter than the actinides suggests that it is impractical to design a clean fission nuclear reactor that uses a static magnetic field as a fission catalyst. However, the situation is not hopeless because it is shown in the following sections that it may be possible to catalyze thermal neutron induced clean fission reactions in the lighter elements by using γ rays.

C. Electromagnetic Wave Resonance Catalysis of Thermal Neutron Induced Clean Fission Reactions for Nuclei Lighter than the Actinides.

The large static magnetic fields predicted to be required for clean fission reactions of subactinide nuclei using thermal neutrons can be circumvented by using high frequency time dependent electromagnetic fields. Nature is full of surprises, and it will now be shown that the superfluid nature of nuclear

matter suggests that using electromagnetic waves (γ rays) that are tuned to a critical frequency for each subactinide element to be fissioned requires a magnetic vector component B_Y^F which is of the order of magnitude $B_Y^F \sim 20$ kilotesla which is much smaller in value than the static field value $B^F \sim 40$ teratesla and can be obtained in laboratory γ ray fluxes (Table 2).

Consider first a static field loading on an atomic nucleus. The spring constant of an atomic nucleus can be calculated in terms of the shear modulus, Poisson's ratio and radius of the nuclear matter in a nucleus. A cylinder of nuclear matter of length h and radius R has a spring constant given by^{95,105}

$$k = E\pi R^2/h \quad (614)$$

where k = spring constant, E = Young's modulus, R = radius of cylinder and h = length of cylinder. The Young's modulus can be expressed in terms of the shear modulus and Poisson's ratio as follows⁹⁵

$$E = 2(1 + \nu)G \quad (615)$$

where ν = Poisson's ratio. Now consider a cylinder whose length is taken according to the condition that the volume of the cylinder is equal to the volume of a spherical nucleus of the same radius

$$\pi R^2 h = 4/3 \pi R^3 \quad (616)$$

which gives

$$h = 4/3R \quad (617)$$

Combining equations (614), (615) and (617) gives the spring constant of an atomic nucleus as

$$k = 3/2\pi(1 + \nu)GR \quad (618)$$

Combining equations (282) and (618) gives

$$k = 3/2\pi b(1 + \nu)GA^{1/3} \quad (619)$$

where the effect of the internal phase angle θ_a is neglected in equation (285), and where b = constant given in equation (282). Values of k for various nuclei appear in Table 2.

For a damped vibrating spring the impedance is given by

$$S = [(k - M_{\text{eff}}\omega^2)^2 + C^2\omega^2]^{1/2} \quad (620)$$

where S = impedance, ω = circular frequency of vibration of a nucleus, M_{eff} = effective dynamic mass of a nucleus, and C = damping constant. The effective mass of a nucleus is given by

$$M_{\text{eff}} = 1/3m_{\text{av}}A = 1/3(Zm_p + Nm_n - B/c^2) \quad (621)$$

where the factor 1/3 enters equation (621) because the mass is distributed throughout the nucleus so that the spring itself is massive.¹⁰⁵ The average nucleon mass that appears in equation (621) is given by

$$m_{av} = Z/Am_p + N/Am_n - \epsilon/c^2 \quad (622)$$

$$\sim 1/2(m_p + m_n) - \epsilon/c^2$$

$$\sim 1/2(m_p + m_n) \sim m_p \sim m_n \quad (623)$$

where c = light speed, ϵ = average binding energy per nucleon given by equation (291), m_p = proton mass and m_n = neutron mass. The dynamics calculation done in this section neglects the effects of the binding energy on the nuclear mass and uses equation (623) for the average mass per nucleon. The approximation that m_{av} = constant makes the effective mass in equation (621) a linear function of the atomic mass number and leads to simple approximate formulas for the resonance frequency and energy. The damping constant that appears in equation (620) is given by

$$C = \eta R = \eta b A^{1/3} \quad (624)$$

where η = viscosity of nuclear matter. The value of the viscosity of nuclear matter is obtained by assuming that it has the same value as the viscosity of superfluid ³He namely $\eta = 2 \times 10^{-7}$ kg/(m sec).¹⁰⁶ The dimensions of the physical quantities that appear in equations (614) through (624) are given by

$$[R] = m$$

$$[b] = m$$

$$[G] = N/m^2 = kg/(m \text{ sec}^2)$$

$$[k] = N/m = kg/\text{sec}^2$$

$$[S] = N/m = kg/\text{sec}^2$$

$$[M_{eff}] = [m_{av}] = [m_p] = [m_n] = kg$$

$$[C] = N \text{ sec}/m = kg/\text{sec}$$

$$[\eta] = kg/(m \text{ sec})$$

$$[\omega] = \text{Hz} = \text{sec}^{-1}$$

The impedance function given by equation (620) has a local minimum value at a frequency which corresponds to the maximum vibration amplitude, and satisfies $S \rightarrow k$ when $\omega \rightarrow 0$.

The simplest way of incorporating the dynamical response of a nucleus to electromagnetic waves into the formalism given by equation (587) for calculating the magnetic induction field required for the clean fission of the subactinides

is to introduce an effective dynamic shear modulus so that in the presence of electromagnetic waves the definition of a dynamic magnetic stiffness modulus for an atomic nucleus is given in analogy to equations (582) and (586) by

$$1/2 G_Y^{\gamma} \theta_z^2 = B_Y'^2 / (2\mu) + 1/2 \epsilon E_Y'^2 = B_Y^2 / (2\mu) = (K_{\theta z}^{BY})^2 \theta_z^2 / (2\mu) \quad (625A)$$

$$K_{\theta z}^{BY} = (\mu G_Y)^{1/2} \quad \text{tesla/deg} \quad (625B)$$

where G_Y = effective dynamic shear modulus which is obtained from equations (618) and (620) as

$$G_Y = 2/(3\pi)(1 + \nu)^{-1} S/R \quad N/m^2 \quad (626)$$

In the limit of $\omega \rightarrow 0$ it follows from equations (620) and (626) that $G_Y \rightarrow G$, so that G_Y is an effective dynamic shear modulus which has the correct static limit. The ratio of interest for clean fission nuclear reactor design is therefore

$$\zeta = K_{\theta z}^{BY} / K_{\theta z}^B = (G_Y/G)^{1/2} = B_Y^F / B^F = [(B_Y^{F'})^2 + \epsilon \mu (E_Y^{F'})^2]^{1/2} / B^F \quad (627)$$

where G_Y is a function of several parameters $G_Y = G_Y(G, \nu, \omega, C, b, A)$. At low frequencies $\zeta \rightarrow 1$. Here $B_Y^{F'} \sim B_Y^F / \sqrt{2}$ and $E_Y^{F'} \sim E_Y^F / \sqrt{2} = B_Y^F / (2\epsilon \mu)^{1/2}$ are the radiation field components required for fission.

The impedance function S for an atomic nucleus has a frequency dependence shown in Figure 2. The ratio ζ of the magnetic stiffness coefficients given in equation (627) also has this frequency dependence and is shown in Figure 3. The impedance S has a minimum value at the giant dipole resonance frequency, and at this frequency the dynamic magnetic stiffness coefficient $K_{\theta z}^{BY}$ and the ratio ζ have much smaller values than their corresponding static values. The reason for this can be discerned from the impedance equation (620) which shows that the giant dipole resonance frequency is given by¹⁰⁵

$$\omega_r = \alpha_c (k/M_{\text{eff}})^{1/2} \quad f_r = \omega_r / (2\pi) \quad (628)$$

where α_c = model correction factor which is introduced to account for the fact that a nucleus is not really described by a simple linear spring model. In this paper the correction factor is chosen to have the value $\alpha_c = 0.5412$, but a different choice for the value of the shear modulus of nuclear matter will yield a different value for α_c in order to agree with measured giant dipole resonance frequencies. Combining equations (619), (620) and (628) gives

$$f_r = \alpha_c (2\pi)^{-1} [9/2\pi b(1 + \nu)G/m_{\text{av}}]^{1/2} A^{-1/3} H_z \quad (629)$$

where m_{av} is given by equation (623). For atomic nuclei the giant dipole resonant frequency is in the γ ray region of the electromagnetic spectrum. The value of the impedance at the giant dipole resonance frequency is obtained from equations (620), (624) and (629) to be

$$\begin{aligned} S_r &= C\omega_r = \alpha_c C(k/M_{\text{eff}})^{1/2} \quad N/m \\ &= \alpha_c nb[9/2\pi b(1 + \nu)G/m_{\text{av}}]^{1/2} \end{aligned} \quad (630)$$

and is seen to be essentially independent of the atomic mass number A because equation (623) shows that m_{av} is roughly independent of A . The damping constant C given by equation (624) was used to obtain the result in equation (630). The value of the dynamic shear modulus at the giant dipole resonance frequency is obtained from equations (282), (626) and (630) to be

$$\begin{aligned} G_r^\gamma &= 2/(3\pi)(1 + \nu)^{-1} S_r / R \quad N/m^2 \\ &= 2/(3\pi)(1 + \nu)^{-1} \alpha_c \eta [9/2\pi b(1 + \nu)G/m_{av}]^{1/2} A^{-1/3} \\ &= \sigma^2 G A^{-1/3} \end{aligned} \quad (631)$$

where the dimensionless number σ is essentially independent of A and is given by

$$\begin{aligned} \sigma &= (G_r^\gamma A^{1/3}/G)^{1/2} \\ &= [2/(3\pi)(1 + \nu)^{-1} S_r / (Gb)]^{1/2} \\ &= \{2b\eta^2 \alpha_c^2 / [\pi m_{av} G(1 + \nu)]\}^{1/4} \\ &= [4b\eta^2 \alpha_c^2 / (\pi m_{av} E)]^{1/4} \end{aligned} \quad (632)$$

where b is defined in equation (282). The dimensionless parameter σ has the value $\sigma = 1.018 \times 10^{-9}$. Then equation (631) gives

$$G_r^\gamma / G = \sigma^2 A^{-1/3} = S_r / k \quad (633)$$

and combining equations (627) and (633) gives the following relationship which is valid at the giant dipole resonance frequency

$$\zeta_r = [K_{\theta z}^{BY} / K_{\theta z}^B]_r = (G_r^\gamma / G)^{1/2} = (S_r / k)^{1/2} = \sigma A^{-1/6} \quad (634)$$

The values of ζ_r for various nuclei appear in Table 2. From equation (627) it follows that the dynamic magnetic stiffness coefficient at the giant dipole resonance frequency is given by

$$\begin{aligned} K_{\theta z}^{BY} &= \zeta_r K_{\theta z}^B \quad \text{tesla/deg} \\ &= \sigma A^{-1/6} K_{\theta z}^B \end{aligned} \quad (635)$$

The dimensionless number σ is still a function of G . The values of $K_{\theta z}^{BY}$ at the giant dipole resonance frequency appear in Table 2 for various atomic nuclei, and Table 2 shows that the values of $K_{\theta z}^{BY}$ are significantly smaller than the value of $K_{\theta z}^B$ given in equation (612).

For electromagnetic waves that are tuned to the γ ray frequencies of the giant dipole resonance frequencies of the subactinide nuclei, the required mag-

netic induction field for the clean fission of subactinide nuclei using thermal neutrons is calculated in analogy to equation (587) as

$$\begin{aligned} B_Y^F &= K_{\theta z}^{BY} \theta_z^F = (\mu G_r^Y)^{1/2} \theta_z^F \quad \text{tesla} \\ &= \sigma A^{-1/6} K_{\theta z}^B \theta_z^F = \sigma A^{-1/6} (\mu G)^{1/2} \theta_z^F \\ &= \sigma A^{-1/6} B^F = \zeta_r B^F \end{aligned} \quad (636)$$

where ζ_r is given by equation (634). The values of B_Y^F for selected nuclei are given in Table 1 and are seen to be much smaller than the values of B^F which also appear in Table 1. The corresponding magnetic field strength of the resonance electromagnetic wave required for clean fission is given by

$$\begin{aligned} H_Y^F &= B_Y^F / \mu = (G_r^Y / \mu)^{1/2} \theta_z^F = \sigma A^{-1/6} H^F \quad \text{amp/m} \\ &= [(H_Y^{F'})^2 + \epsilon / \mu (E_Y^{F'})^2]^{1/2} \end{aligned} \quad (637)$$

where $\mu \sim \mu_0$ = magnetic permeability of nuclear matter. The corresponding values of the electric field of the resonance electromagnetic waves required for the clean fission of subactinide nuclei are given by^{91,92}

$$E_Y^F = (\mu_0 / \epsilon_0)^{1/2} H_Y^F = 120\pi H_Y^F \quad \text{volts/m} \quad (638)$$

where $H_Y^{F'} \sim H_Y^F / \sqrt{2} = B_Y^F / (\sqrt{2} \mu)$ and $E_Y^{F'} \sim E_Y^F / \sqrt{2}$ give the radiation field components. The power density of the resonance electromagnetic waves that are required for the clean fission of subactinide nuclei using thermal neutrons is^{91,92}

$$\begin{aligned} P_Y^F &= 1/2 (\mu_0 / \epsilon_0)^{1/2} (H_Y^F)^2 \quad \text{W/m}^2 \\ &= 1/2 (120\pi) (B_Y^F / \mu_0)^2 \\ &= 1/2 (120\pi) (\sigma^2 G / \mu_0) A^{-1/3} (\theta_z^F)^2 \end{aligned} \quad (639)$$

where θ_z^F = critical internal angle of the complex atomic number given by equation (483) and corresponding to incipient fission of a subactinide nucleus. The γ photon energy required to catalyze clean fission of the subactinides can be obtained from equation (629) to be

$$\epsilon_Y^F = h\nu_Y^F = hf_r \quad \text{MeV/photon} \quad (640)$$

where h = Planck's constant and has the value¹⁰⁶

$$\begin{aligned} h &= 6.626 \times 10^{-34} \quad \text{Joule sec} \\ &= 4.136 \times 10^{-21} \quad \text{MeV sec} \end{aligned} \quad (641)$$

The γ ray flux density required to catalyze clean fission of subactinide nuclei using thermal neutrons is obtained from the γ ray power density given in equation (639) as follows

$$\phi_Y^F = P_Y^F / \epsilon_Y^F \quad \text{photons}/(\text{sec m}^2) \quad (642)$$

The γ ray photon number density required for clean fission is obtained from the flux density given in equation (642) by

$$n_Y^F = \phi_Y^F / c \quad \text{photons/m}^3 \quad (643)$$

where c = light speed given by¹⁰⁶

$$c = 2.9979 \times 10^8 \text{ m/sec} \quad (644)$$

The quantities H_Y^F , E_Y^F , P_Y^F , ϵ_Y^F , ϕ_Y^F and n_Y^F evaluated at the giant dipole resonance frequency appear in Table 3.

The mechanical and electromagnetic quantities that are relevant to the clean fission of subactinide nuclei depend of the value chosen for Poisson's ratio as for example in equation (632). The value of Poisson's ratio for nuclear matter is not known, but it is generally in the range of values $0 < \nu < 0.5$.⁹⁵ Negative values of Poisson's ratio are possible but only in very exotic materials and will not be considered here. Because the physical quantities calculated in this section are not particularly sensitive to the choice of value for Poisson's ratio, the choice $\nu = 1/4$ is made for the numerical calculations. Then the elastic properties of nuclear matter are taken from equation (604) to be

$$G = 3.98 \times 10^{33} \text{ kg}/(\text{m sec}^2) \quad \nu = 1/4$$

Nuclear matter is a viscous superfluid whose viscosity is assumed to have the same value as does superfluid liquid ^3He .¹⁰⁶ Then the following values of nuclear properties are used to produce Tables 1 through 3:

$$m_{av} = 1.6738 \times 10^{-27} \text{ kg} \quad (646)$$

$$b = 1.2 \text{ fm} \quad (647)$$

$$\eta = 2.0 \times 10^{-7} \text{ kg m}^{-1} \text{ sec}^{-1} \quad (648)$$

$$\sigma = 1.018 \times 10^{-9} \quad (649)$$

$$S_r = 29.17 \text{ N m}^{-1} \quad (650)$$

$$k = 28.10 \times 10^{18} \text{ A}^{1/3} \text{ N m}^{-1} \quad (651)$$

$$f_r = 19.34 \times 10^{21} \text{ A}^{-1/3} \text{ Hz} \quad (652)$$

$$\lambda_r = 15.50 \text{ A}^{1/3} \text{ fm} \quad (653)$$

$$\epsilon_Y^F = 80.0 \text{ A}^{-1/3} \text{ MeV/photon} \quad (654)$$

Values of k , f_r and λ_r appear in Table 2, while ϵ_Y^F appears in Table 3. The γ ray photon number density n_Y^F , flux density ϕ_Y^F and power density P_Y^F required

for fission depend on Z and A through the fission angle $\theta_Z^F(Z,A)$ as shown in equation (639). The value of the impedance at resonance S_r is independent of A . Note that k is 18 orders of magnitude larger than S_r which corresponds to the fact that B^F is 9 orders of magnitude larger than B_Y^F as seen from Table 1. The relatively low values of S_r and B_Y^F is due to the fact that at the resonance frequency it is the small value of the superfluid viscosity of nuclear matter that determines the value of the impedance.

Table 1 gives Z^2/A , the fissility parameter χ , the internal phase angle of the proton number θ_Z^F required for the clean fission of nuclei by thermal neutrons, the static magnetic induction field B^F required to catalyze clean fission of nuclei with thermal neutrons, and the dynamic magnetic induction field B_Y^F of γ rays at the giant dipole resonance frequency which is required to catalyze clean fission of nuclei by thermal neutrons. Table 2 gives the ratio ζ_r of the dynamic magnetic induction field to the static magnetic induction field required for clean fission, the dynamic magnetic stiffness coefficient $K_{\theta_Z}^{BY}$ for γ rays at the giant dipole resonance frequency, the nuclear spring constant k , the giant dipole resonance frequency f_r , and the wavelength λ_r of γ rays corresponding to the giant dipole resonance frequency. Table 3 gives the dynamic magnetic field H_Y^F of γ rays at the giant dipole resonance frequency which is required to catalyze clean fission in atomic nuclei using thermal neutrons, the dynamic electric field strength E_Y^F of γ rays at the giant dipole frequency required to catalyze clean fission in nuclei using thermal neutrons, the resonant power density P_Y^F of γ rays required for the clean fission of nuclei utilizing low energy thermal neutrons, the resonant γ ray energy ϵ_Y^F required for catalysis of clean fission, the resonant γ ray flux ϕ_Y^F required for clean fission catalysis, and the resonant γ ray photon number density n_Y^F required to catalyze clean fission in atomic nuclei using low energy thermal neutrons.

For clean fission nuclear reactions of the subactinide nuclei using thermal neutrons, the γ rays are used only to bring the internal phase angle of the atomic number θ_Z up to its critical value θ_Z^F required for fission as described in Section 5. The very small value of the viscosity of nuclear matter gives a small value to the parameter σ that is defined for the resonance frequency condition by equations (631), (632) and (649). The small value of σ when used in equations (636) through (639) gives the relatively low values for the electric and magnetic fields, power density and flux of the resonant γ rays that are required for the clean fission of subactinide nuclei. Therefore immersing subactinide nuclei in a bath of γ rays that are tuned to the giant resonance frequency of the nuclei requires only a relatively low power density of γ rays to catalyze clean fission using thermal neutrons. The low γ ray intensity (associated with G_r^Y and S_r of the nuclei) required for clean fission at the giant dipole resonance frequency compared to the high values of the static magnetic field (associated with G and k of the nuclei) required for fission in a static field is related to the behavior of the nuclear impedance S given by equation (620) and shown in Figure 2 which indicates that the impedance has a deep minimum at the giant dipole resonance frequency. The γ ray intensity required for clean fission has a minimum value when the energy of the incident γ rays corresponds to the giant dipole resonance frequency of the subactinide element that is chosen as a nuclear fuel. Thus thermal neutron induced fission in the subac-

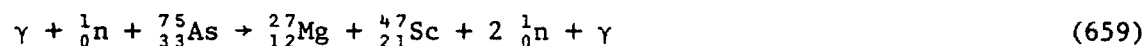
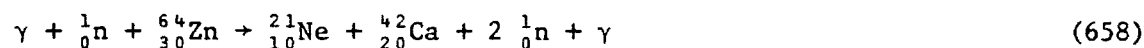
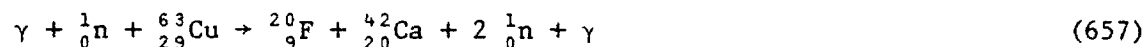
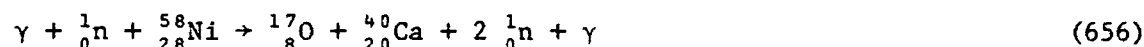
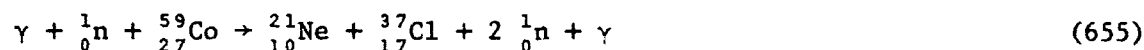
tinides is catalyzed in the simplest way by γ rays tuned to the giant dipole resonance frequency of nuclei as given by equations (652) and (653), and whose intensity is determined by the critical condition $\theta_z = \theta_z^F$ as in equations (483) and (636) through (639).

D. γ Ray Catalyzed Low Energy Clean Fission Nuclear Reactions.

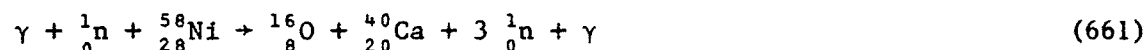
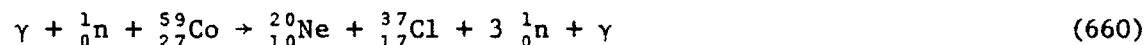
Several examples of γ ray catalyzed clean fission reactions induced by thermal neutrons are now considered. The average incident thermal neutron energy is 0.025 MeV.

Intermediate Weight Nuclei.

Examples of clean fission reactions whose fission products are low level beta emitters are



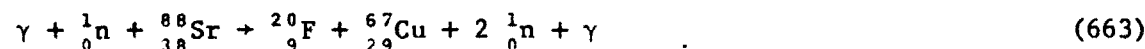
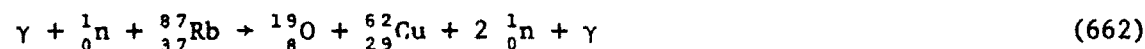
These fission reactions produce low level radionuclides as waste products. Examples of perfectly clean fission reactions using thermal neutrons and γ ray catalysis are

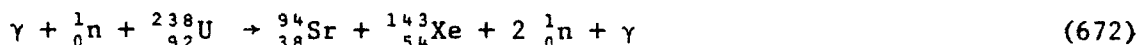
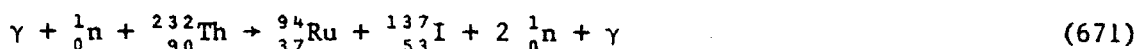
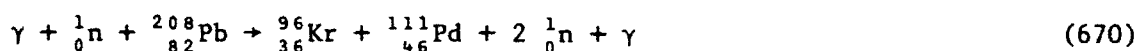
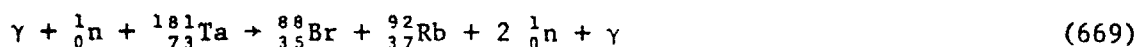
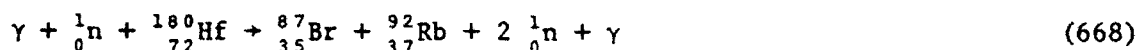
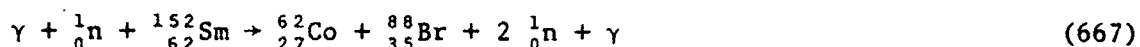
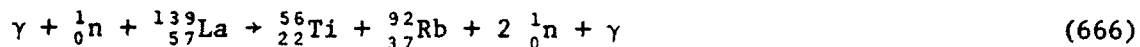
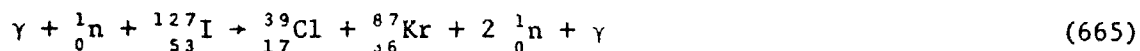
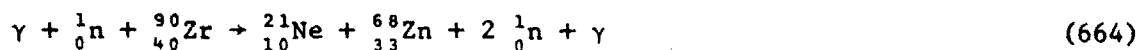


The fission products of these reactions are not radioactive and can be used for the development of clean fission nuclear reactors.

Heavy Nuclei.

The following are examples of γ ray catalyzed fission reactions using thermal neutrons whose fission products can be relatively low level radionuclides as in the case when ${}^{87}\text{Rb}$ and ${}^{88}\text{Sr}$ are fissioned, but with high level radionuclides as fission products when the heavy elements are fissioned such as in the case of ${}^{238}\text{U}$ and ${}^{232}\text{Th}$





These fission reactions would not occur using thermal neutrons without the catalytic effects of the resonant γ rays.

Nuclear fission is a statistical process of penetration through a fission barrier, and many other fission reactions are possible. For a sustained fission chain reaction at least two neutrons must be produced by the fission reactions. For both intermediate and heavy nuclei generally at least one of the fission product nuclei of these reactions is beta unstable for clean fission reactions. Equations (660) and (661) are exceptional in that the fission products are not radioactive. Because the fission product nuclei of intermediate weight elements are relatively light weight nuclei close to the valley of beta stability, the intensity of beta, alpha and neutron radioactivity of the fission products is low level for the γ ray catalyzed fission of intermediate weight elements using thermal neutrons as, for example, in equations (655) through (659). Equations (662) through (672) show that the γ ray catalyzed fission reactions of heavy nuclei using thermal neutrons yield fission products that are relatively high level beta, alpha and neutron emitters because the fission products generally have large neutron excesses. Therefore for the development of clean fission nuclear reactors fuel elements only slightly heavier than ${}^{26}\text{Fe}$ must be used as shown in equations (655) through (661).

E. Photonuclear Reactions and Clean Fission for Subactinide Nuclei.

This section considers the form of the electric dipole sum rule for γ ray catalyzed thermal neutron induced fission reactions in nuclei whose fissility parameters satisfy $\chi \leq 1$. Photonuclear reactions are induced by γ ray photons interacting with atomic nuclei. For low energy photons the process in its simplest form is due to resonance fluorescence which occurs when photon absorption by a nucleus is followed by photon emission as the nucleus decays to the ground state. Nucleon emission and nuclear fission can occur for higher energies of the incident γ rays. At photon energies of about 12 - 20 MeV the giant dipole resonance is excited.^{60-67,108}

The conventional electric dipole sum rule for photonuclear reactions is written in the standard incoherent spacetime form as follows^{60-67,108}

$$G_{inc} = \int \sigma d\epsilon = gZN/A \quad (673)$$

where σ = photonuclear reaction cross section for incoherent spacetime, Z , N and A = atomic number, neutron number and atomic mass number of the target nucleus, and where¹⁰⁸

$$g = 2\pi^2 e^2 \hbar / (m_{av} c) \sim 0.06 \text{ MeV b} \quad (674)$$

where the integral is taken over photon energies up to 30 MeV. The concept of the broken symmetry forms of the atomic number, neutron number and atomic mass number suggests that a complex number generalization of the photonuclear reaction sum rule should be written as

$$\bar{G} = \int \bar{\sigma} d\bar{\epsilon} = g\bar{z}\bar{n}/\bar{a} \quad (675)$$

where $\bar{\sigma}$ = complex number photonuclear reaction cross section, \bar{G} = complex number integrated photonuclear cross section, and \bar{z} , \bar{n} and \bar{a} = complex number atomic number, neutron number and atomic mass number respectively which are given by equations (40) through (42). The complex numbers $\bar{\sigma}$ and \bar{G} can be represented as

$$\bar{\sigma} = \sigma \exp(j\theta_\sigma) \quad \bar{G} = G \exp(j\theta_G) \quad (676)$$

which are complex numbers in an internal space. Equation (675) can be written as

$$G = gzn/a = gZN/A \cos \theta_z \cos \theta_n \cos^{-1} \theta_a \quad (677)$$

$$\theta_G = \theta_z + \theta_n - \theta_a \quad (678)$$

Equation (675) can also be written as

$$\bar{G} = \int_0^\infty \sigma \sec \beta_{\epsilon\epsilon} \exp[j(\theta_\sigma + \theta_\epsilon + \beta_{\epsilon\epsilon})] d\epsilon \quad (679)$$

$$= \int_0^{\pi/6} \sigma \epsilon \csc \beta_{\epsilon\epsilon} \exp[j(\theta_\sigma + \theta_\epsilon + \beta_{\epsilon\epsilon})] d\theta_\epsilon \quad (680)$$

where the complex number photon energy is written as

$$\bar{\epsilon} = \epsilon \exp(j\theta_\epsilon) \quad (681)$$

and where

$$\tan \beta_{\epsilon\epsilon} = \epsilon \partial \theta_\epsilon / \partial \epsilon \quad (682)$$

The component form of equations (679) and (680) are written as

$$G \cos \theta_G = \int_0^{\infty} \sigma \sec \beta_{\epsilon\epsilon} \cos(\theta_\sigma + \theta_\epsilon + \beta_{\epsilon\epsilon}) d\epsilon \quad (683)$$

$$= \int_0^{\pi/6} \sigma \epsilon \csc \beta_{\epsilon\epsilon} \cos(\theta_\sigma + \theta_\epsilon + \beta_{\epsilon\epsilon}) d\theta_\epsilon \quad (684)$$

$$G \sin \theta_G = \int_0^{\infty} \sigma \sec \beta_{\epsilon\epsilon} \sin(\theta_\sigma + \theta_\epsilon + \beta_{\epsilon\epsilon}) d\epsilon \quad (685)$$

$$= \int_0^{\pi/6} \sigma \epsilon \csc \beta_{\epsilon\epsilon} \sin(\theta_\sigma + \theta_\epsilon + \beta_{\epsilon\epsilon}) d\theta_\epsilon \quad (686)$$

which are generally valid equations. From equations (677) and (678) it follows that

$$G \cos \theta_G = gZN/A \cos \theta_z \cos \theta_n \cos^{-1} \theta_a \cos(\theta_z + \theta_n - \theta_a) \quad (687)$$

$$G \sin \theta_G = gZN/A \cos \theta_z \cos \theta_n \cos^{-1} \theta_a \sin(\theta_z + \theta_n - \theta_a) \quad (688)$$

Generally $G \cos \theta_G > 0$ if $Z > 0$, $N > 0$ and $A > 0$.

The upper integration limit of $\pi/6$ in equations (680), (684) and (686) arises from the conservation of momentum for the photon-nucleon interaction which can be written as

$$\bar{\epsilon}/c = h\bar{\nu}/c = m\bar{v} \quad (689)$$

where the complex number photon frequency $\bar{\nu}$ and the complex number nucleon velocity \bar{v} are written as

$$\bar{\nu} = \nu \exp(j\theta_\nu) \quad \bar{v} = v \exp(j\theta_v) \quad (690)$$

Equations (689) and (690) can be written as

$$\epsilon/c = h\nu/c = mv \quad \theta_\epsilon = \theta_\nu = \theta_v = \theta_x - \theta_t \quad (691)$$

which are valid for the inelastic photonuclear reaction. For coherent spacetime $\theta_x = \pi/3$, $\theta_t = \pi/6$ and $\theta_v = \pi/6$, so that $\theta_\epsilon = \pi/6$ for inelastic photonuclear reactions in coherent spacetime. It should be noted that for elastic collisions of photons with particles, the phase angle relation is $\theta_\epsilon = 2\theta_v$, with the result that $\theta_\epsilon = \pi/3$ for elastic collisions in coherent spacetime.

If the photon energy is taken to be incoherent it can be represented as a real number with $\theta_\epsilon = 0$ and $\beta_{\epsilon\epsilon} = 0$ and equation (675) becomes

$$\bar{G} = \int \bar{\sigma} d\epsilon = g\bar{z}\bar{n}/\bar{a} \quad (692)$$

while equations (683) and (685) give

$$G \cos \theta_G = \int_0^{\infty} \sigma \cos \theta_{\sigma} d\epsilon \quad (693)$$

$$G \sin \theta_G = \int_0^{\infty} \sigma \sin \theta_{\sigma} d\epsilon \quad (694)$$

which are valid for incoherent photon energies. The lefthand side of equations (693) and (694) are given by equations (687) and (688). If the photon energy is taken to be coherent with $d\bar{\epsilon} = j\bar{\epsilon}d\theta_{\epsilon}$ where the photon energy magnitude is given by $\epsilon = \epsilon_c = \text{constant}$ and where $\theta_{\epsilon} = \pi/2$, then equation (675) becomes

$$\bar{G} = j \int_0^{\pi/6} \bar{\sigma} \bar{\epsilon} d\theta_{\epsilon} = g\bar{Z}\bar{n}/\bar{A} \quad (695)$$

and equations (684) and (686) become

$$G \cos \theta_G = -\epsilon_c \int_0^{\pi/6} \sigma \sin(\theta_{\sigma} + \theta_{\epsilon}) d\theta_{\epsilon} \quad (696)$$

$$G \sin \theta_G = \epsilon_c \int_0^{\pi/6} \sigma \cos(\theta_{\sigma} + \theta_{\epsilon}) d\theta_{\epsilon} \quad (697)$$

which are valid for coherent photon energies. If $\bar{\sigma}$ is independent of θ_{ϵ} then equations (696) and (697) become with $\sigma = \sigma_c = \text{constant}$ for coherent photon energies

$$G \cos \theta_G = \epsilon_c \sigma_c [\cos(\theta_{\sigma} + \pi/6) - \cos \theta_{\sigma}] \quad (698)$$

$$G \sin \theta_G = \epsilon_c \sigma_c [\sin(\theta_{\sigma} + \pi/6) - \sin \theta_{\sigma}] \quad (699)$$

Therefore in general $\theta_G \geq \pi/2$ for coherent photon energies, and in particular if $\theta_{\sigma} = 0$ then $\theta_G = 105^\circ$. This means that $G \cos \theta_G < 0$ for coherent photon energies. Therefore a coherent spacetime photonuclear reaction that is described by the sum rule in equation (695) is not possible unless $Z < 0$, $N < 0$ and $A < 0$ in equation (687) which is a case analogous to the negative values of the ordinary magnetic quantum number m .

For light nuclei in the valley of beta stability equation (178) gives

$$\theta_z \sim \theta_n \sim \theta_a \quad (700)$$

and for this special case equation (677) and (678) become

$$G = gZN/A \cos \theta_a \quad (701)$$

$$\theta_G = \theta_a \quad (702)$$

For this case equations (693) and (694) become for incoherent photon energies

$$gZN/A \cos^2 \theta_a = \int \sigma \cos \theta_{\sigma} d\epsilon \quad (703)$$

$$gZN/A \cos \theta_a \sin \theta_a = \int \sigma \sin \theta_{\sigma} d\epsilon \quad (704)$$

$$\tan \theta_a = (\int \sigma \sin \theta_{\sigma} d\epsilon) / (\int \sigma \cos \theta_{\sigma} d\epsilon) \quad (704A)$$

from which σ and θ_σ can be obtained by inversion techniques for nuclei in the valley of beta stability.

Consider now the case of nuclei with fissility parameters less than unity $\chi \leq 1$ that have been brought to the condition of incipient fission by the application of an external γ ray field. For this case the internal phase angles of the atomic number, neutron number and atomic mass number are related by equation (473) as follows

$$\theta_a^F = 2\theta_z^F = 2\theta_n^F \quad (705)$$

and therefore equations (677), (678) and (705) give

$$G = gZN/A(1 - \tan^2 \theta_z^F)^{-1} \quad (706)$$

$$\theta_G = 0 \quad (707)$$

From equations (685) and (707) it follows that for incipient fission

$$\theta_\sigma = 0 \quad \theta_\epsilon = 0 \quad \beta_{\epsilon\epsilon} = 0 \quad (708)$$

so that the photonuclear interaction for the case of fission must be scalar ($\theta_\sigma = 0$) with incoherent photon interactions, and equations (693), (694) and (706) through (708) become

$$\int_0^\infty \sigma d\epsilon = gZN/A(1 - \tan^2 \theta_z^F)^{-1} \quad (709)$$

However it has been shown in equation (480) that the approximate condition for γ ray catalyzed thermal neutron induced clean fission is given by

$$\tan \theta_z^F = (1 - \chi)^{1/2} \quad (710)$$

where $\chi \leq 1$ is the fissility parameter. Combining equations (709) and (710) gives for incoherent photon interactions

$$\int_0^\infty \sigma d\epsilon = g\chi^{-1}ZN/A \quad (711)$$

and therefore the fissility parameter enters into the electric dipole sum rule for γ ray catalyzed thermal neutron induced clean fission reactions.

8. γ RAY CATALYZED CLEAN FISSION NUCLEAR REACTOR DESIGN CONCEPTS. This section presents design concepts for γ ray catalyzed clean fission nuclear reactor cores whose fuel consists of elements heavier than ^{56}Fe but lighter than the actinides. As explained in the previous sections the subactinide elements cannot be fissioned by thermal neutrons under ordinary circumstances, but in the presence of γ rays tuned to the giant resonance frequency of the fuel elements these nuclei can be fissioned by thermal neutrons. The γ ray photon number density required to fission a subactinide nucleus is relatively low because the γ rays are required only to bring the internal phase angle θ_z of the

atomic number up to its incipient fission value θ_Z^F given by equation (483) at which point fission by thermal neutrons is possible. When elements heavier than ^{56}Fe , such as for example ^{63}Cu , are used as fuel in a clean fission nuclear reactor, the fission products of the thermal neutron induced γ ray catalyzed fission reactions are neutrons and atomic nuclei which have relatively low level radioactivity because they have smaller neutron excesses than those of the fission product nuclei of ^{235}U or ^{239}Pu which are the fuel elements of conventional nuclear reactors.

Relatively high energy γ rays with energies in the range of 10-20 MeV are required for the clean fission of subactinide nuclei. Possible sources of the relatively high energy γ rays that are required to catalyze clean fission reactions are

- a) bremsstrahlung of electrons in matter
- b) nuclei excited by collisions and subsequent giant dipole resonance decay
- c) synchrotron radiation from electrons
- d) nuclear reactions
- e) nuclear fission
- f) γ ray lasers

The possibility of using γ rays from the radioactive decay of natural and man-made radioisotopes is excluded because they generally are in the range $0.01 < \epsilon_\gamma < 10$ MeV which is too low to excite the giant dipole resonance states in atomic nuclei.

Figures 4 and 5 show two clean fission nuclear reactor core design concepts in which γ rays are directed to the surface of a fuel and heat exchanger system. In this case the fuel element, such as ^{63}Cu , is in the form of rectangular plates or a spherical shell which are assembled adjacent to a structure which contains a heat exchanger. Table 3 shows that for ^{63}Cu as the fuel element the incident γ rays must be in the range of $\epsilon_\gamma = 20$ MeV and have a photon number density of at least $n_\gamma = 7 \times 10^{23}$ photons/ m^3 which is a relatively small number density. The γ rays impinge on the surfaces of the fuel element plates and elevate the value of the internal phase angle θ_Z of the atomic number of the fuel element up to the critical value θ_Z^F required for clean fission which for ^{63}Cu is $\theta_Z^F = 38.5^\circ$ as shown in Table 1. Then an initial source of thermal neutrons is introduced to start the fission reaction. The fission reaction emits two or three high energy neutrons as a fission product and these can be thermalized by moderators to produce a self sustaining chain reaction. The clean fission nuclear reactions are expected to occur mainly at the surface of the fuel element plates because γ rays attenuate rapidly in matter.

After an initial introduction of thermal neutrons and an initial input of power to activate the γ ray sources the output of neutrons and power from the clean fission reactions can be used to create a self sustaining nuclear reactor. During operation of the reactor a portion of the power developed is used to drive the external γ ray sources. The nuclear reaction rates can be controlled by adjusting the external γ ray flux, so that the control rods of the standard uranium fission reactors are not required to control the neutron flux in clean fission reactors but can be retained as a safety factor. Minimum shielding is required for a clean fission nuclear reactor because of the

low level radioactivity of the fission product nuclei. Minimum shielding is also required for high energy γ rays because they are rapidly attenuated in matter by electron-positron pair production.

9. CONCLUSION. This paper suggests that in the presence of an electromagnetic or gravitational field the atomic number, neutron number and atomic mass number have broken internal symmetries and must be represented as complex numbers of a special type. The magnitudes and phase angles of these complex numbers are determined by the requirement that the nuclear wave function be periodic in terms of the measured values of the internal angle coordinates associated with the atomic number, neutron number and atomic mass number. In the presence of an external field the number of atomic nuclei must also be represented as a complex number, and this suggests the possibility of internal phase angle radioactive decays wherein the integer numbers Z , N , A and η remain fixed but their associated internal phase angles are decaying with time. For all types of nuclear reactions and decays the integer baryon number is conserved.

A complex number form of the liquid drop nuclear mass formula can be developed and applied to the calculation of the binding energy of atomic nuclei. Similarly, a complex number Bohr-Wheeler fission equation can be developed that describes the condition for spontaneous and thermal neutron induced fission of atomic nuclei located in an electromagnetic or gravitational field. This fission condition suggests that the internal phase angles associated with Z , N and A can be affected by the application of an electromagnetic field in such a way that nuclear fission by thermal neutrons can be catalyzed by γ rays in subactinide nuclei for which the fissility parameter satisfies $\chi < 1$ and which ordinarily would not undergo fission by thermal neutrons.

The catalysis of thermal neutron induced clean fission requires a) that the frequency of the incident γ rays be tuned to the giant dipole resonance frequency of the particular subactinide element that is selected to be used as fuel in a nuclear reactor, and b) that the incident γ ray photon number density be larger than a critical value required to make the internal phase angle of the atomic number θ_Z equal to a critical value θ_Z^F that is required for fission. This means that γ rays in the range of 12-25 MeV must be used to catalyze thermal neutron induced fission in the subactinide elements, for example ^{63}Cu requires 20 MeV γ rays. The fission of subactinide elements will produce fission products that have low level beta, alpha and neutron emissions, so that the radioactive wastes from clean fission nuclear reactors will be not nearly as dangerous as the fission waste products from conventional ^{235}U or ^{239}Pu nuclear reactors. By using intermediate weight elements such as ^{59}Co , ^{58}Ni or ^{63}Cu as fuel elements for γ ray catalyzed thermal neutron induced fission, it is possible to construct clean fission nuclear reactors that have no radioactive waste products whatsoever. Finally, the dangerously radioactive fission waste product elements of conventional ^{235}U and ^{239}Pu nuclear reactors can be used as fuel in γ ray catalyzed clean fission reactors, thereby offering a useful way of eliminating present day stored radioactive wastes.

ACKNOWLEDGEMENT

This work could not have been completed without the kind help of Elizabeth K. Klein who typed and edited this paper.

REFERENCES

1. Clark, K., Animals and Men, William Morrow, New York, 1977.
2. Krupnick, A. J. and Portney, P. R., "Controlling Urban Air Pollution: A Benefit-Cost Assessment," *Science*, Vol. 252, p. 522, 26 Apr 1991.
3. Corcoran, E., "Cleaning Up Coal," *Scientific American*, p. 107, May 1991.
4. Stern, A. C., editor, Air Pollution, Vols. 1-5, Academic, New York, 1976.
5. Meszaros, E., Atmospheric Chemistry-Fundamental Aspects, Elsevier, New York, 1981.
6. Wark, K. and Warner, C. F., Air Pollution, Harper and Row, New York, 1981.
7. Brimblecombe, P., Air-Composition and Chemistry, Cambridge Univ. Press, New York, 1986.
8. Campbell, I. M., Energy and the Atmosphere, John Wiley, New York, 1986.
9. Pigford, T. H., "Environmental Aspects of Nuclear Energy Production," *Ann. Rev. Nucl. Part. Sci.*, Vol 24, p. 515, 1974.
10. Roberts, L. E. J., "Radioactive Waste Management," *Ann. Rev. Nucl. Part. Sci.*, Vol. 40, p. 79, 1990.
11. Kathren, R. L., Radioactivity in the Environment, Harwood, New York, 1984.
12. Chapman, N. A. and McKinley, I. G., The Geological Disposal of Nuclear Waste, John Wiley, New York, 1987.
13. Berlin, R. E. and Stanton, C. C., Radioactive Waste Management, John Wiley, New York, 1989.
14. Tang, Y. S. and Saling, J. H., Radioactive Waste Management, Hemisphere Publishing Corp., New York, 1990.
15. Krauskopf, K. B., "Disposal of High-Level Nuclear Waste: Is it Possible?," *Science*, Vol. 249, p. 1231, 14 September 1990.
16. Steahan, R. T., Alternative Energy Sources, Aspen Publications, Rockville, MD, 1981.
17. Furlan, G., Mancini, N. A. and Sayigh, A. A. M., Nonconventional Energy, Plenum, New York, 1984.
18. Edmonds, J. and Reilly, J. M., Global Energy, Oxford University Press, New York, 1985.
19. Veziroglu, T. N., editor, Alternative Energy Sources VIII, Vols. 1 and 2, Hemisphere Publishing Corp., New York, 1989.

20. Cassedy, E. S. and Grossman, P. Z., Introduction to Energy, Cambridge Univ. Press, New York, 1990.
21. Helm, J. L., Energy, National Academy Press, Washington, D. C., 1990.
22. Gibbons, J. H. and Blair, P. D., "US Energy Transition: On Getting from Here to There," *Physics Today*, July 1991.
23. Putnam, P. C., Power from the Wind, Van Nostrand Reinhold, New York, 1948.
24. Inglis, D. R., Wind Power, Univ. of Michigan Press, Ann Arbor, 1978.
25. DeRenzo, D. J., Wind Power: Recent Developments, Noyes Data Corp, Park Ridge, New Jersey, 1979.
26. Sayigh, A., editor, Solar Energy Engineering, Academic Press, New York, 1977.
27. Kreith, F. and Kreider, J. F., Principles of Solar Engineering, McGraw-Hill, New York, 1978.
28. Duffie, J. A. and Beckman, W. A., Solar Engineering of Thermal Processes, John Wiley, New York, 1980.
29. Hall, D. O. and Morton, J., editors, Solar World Forum, Vols. 1-4, Pergamon Press, New York, 1982.
30. Berman, E. R., Geothermal Energy, Noyes Data Corp., Park Ridge, New Jersey, 1975.
31. Goguel, J., Geothermics, McGraw-Hill, New York, 1976.
32. Ellis, A. J. and Mahon, W. A. J., Chemistry and Geothermal Systems, Academic Press, New York, 1977.
33. Post, R. F., "Controlled Fusion Research and High-Temperature Plasmas," *Ann. Rev. Nucl. Part. Sci.*, Vol. 20, p. 509, 1970.
34. Hirsch, R. L., "Status and Future Directions of the World Program in Fusion Research and Development," *Ann. Rev. Nucl. Part. Sci.*, Vol. 25, p. 79, 1975.
35. Motz, H., The Physics of Laser Fusion, Academic Press, New York, 1979.
36. Teller, E., editor, Fusion, Vol. 1, Parts A & B, Academic Press, New York, 1981.
37. Keefe, D., "Inertial Confinement Fusion," *Ann. Rev. Nucl. Part. Sci.*, Vol. 32, p. 391, 1982.
38. Dolan, T. J., Fusion Research, Pergamon Press, New York, 1982.

39. Stacey, W. M., Fusion, John Wiley, New York, 1984.
40. Niu, K., Nuclear Fusion, Cambridge Univ. Press, New York, 1989.
41. Breunlich, W. H., Kammel, P., Cohen, J. S. and Leon, M., "Muon-Catalyzed Fusion," Ann. Rev. Nucl. Part. Sci., Vol. 39, p. 311, 1989.
42. Furth, H. P., "Magnetic Confinement Fusion," Science, Vol. 249, p. 1522, 28 Sep. 1990.
43. Friar, J. L., Gibson, B. F., Jean, H. C. and Payne, G. L., "Nuclear Transition Rates in μ -Catalyzed p-d Fusion," Phys. Rev. Lett., Vol. 66, p. 1827, 8 Apr. 1991.
44. Corday, J. G., Goldston, R. J. and Parker, R. R., "Progress Toward a Tokamak Fusion Reactor," Physics Today, p. 22, January 1992.
45. Callen, J. D., Carreras, B. A. and Stambaugh, R. D., "Stability and Transport Processes in Tokamak Plasmas," Physics Today, p. 34, January 1992.
46. Weinberg, A. M. and Wigner, E. P., The Physical Theory of Neutron Chain Reactors, University of Chicago Press, Chicago, 1958.
47. Etherington, H., editor, Nuclear Engineering Handbook, McGraw-Hill, New York, 1958.
48. Glasstone, S. and Sesonske, A., Nuclear Reactor Engineering, Van Nostrand Reinhold, New York, 1967.
49. Hafele, W., Faude, D., Fischer, E. A. and Laue, H. J., "Fast Breeder Reactors," Ann. Rev. Nucl. Part. Sci., Vol. 20, p. 393, 1970.
50. Perry, A. M. and Weinberg, A. M., "Thermal Breeder Reactors," Ann. Rev. Nucl. Part. Sci., Vol. 22, p. 317, 1972.
51. Lamarsh, J. R., Introduction to Nuclear Reactor Theory, Addison-Wesley, New York, 1972.
52. Zweifel, P. F., Reactor Physics, McGraw-Hill, New York, 1973.
53. Lamarsh, J. R., Introduction to Nuclear Engineering, Addison-Wesley, Reading, MA, 1975.
54. Taylor, T. B., "Nuclear Safeguards," Ann. Rev. Nucl. Part. Sci., Vol. 25, p. 407, 1975.
55. Bell, G. I. and Glasstone, S., Nuclear Reactor Theory, Krieger Publishing Co., Huntington, NY, 1979.
56. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.

57. Weiss, R. A., "Electromagnetism and Gravity," Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 91-1, p. 265, June 19-22, 1990.
58. Romer, A., editor, The Discovery of Radioactivity and Transmutation, Dover, New York, 1964.
59. Rasetti, F., Elements of Nuclear Physics, Prentice-Hall, New York, 1936
60. Green, A. E. S., Nuclear Physics, McGraw-Hill, New York, 1955.
61. Evans, R. D., The Atomic Nucleus, McGraw-Hill, New York, 1955.
62. Eder, G., Nuclear Forces, MIT Press, Cambridge, 1968.
63. Elton, L. R. B., Introductory Nuclear Theory, Interscience, New York, 1959.
64. DeBenedetti, S., Nuclear Interactions, John Wiley, New York, 1964.
65. Blatt, J. M. and Weisskopf, V. F., Theoretical Nuclear Physics, John Wiley, New York, 1952.
66. Bethe, H. A. and Morrison, P., Elementary Nuclear Theory, John Wiley, New York, 1961.
67. Greiner, W., Nuclear Theory, Vols. 1-3, Elsevier, New York, 1976.
68. Weiss, R. A. and Cameron, A. G. W., "Equilibrium Theory of the Nuclear Symmetry Energy of Infinite Nuclear Matter," and "Equilibrium Theory of the Symmetry Energy of Finite Nuclei," Can. J. Phys., Vol. 47, p. 2171 and p. 2211, 1969.
69. Willets, L., Theories of Nuclear Fission, Oxford Univ. Press, New York, 1964.
70. Fong, P., Statistical Theory of Nuclear Fission, Gordon & Breach, New York, 1969.
71. Halpern, I., "Three Fragment Fission," Ann. Rev. Nucl. Part. Sci., Vol. 21, p. 245, 1971.
72. Nix, J. R., "Calculation of Fission Barriers for Heavy and Superheavy Nuclei," Ann. Rev. Nucl. Part. Sci., Vol. 22, p. 65, 1972.
73. Bohr, A. and Mottelson, B. R., "The Many Facets of Nuclear Structure," Ann. Rev. Nucl. Part. Sci., Vol. 23, p. 363, 1973.
74. Vandebosch, R. and Huizenga, J. R., Nuclear Fission, Academic Press, New York, 1973.
75. Hoffman, D. C. and Hoffman, M. M., "Post-Fission Phenomena," Ann. Rev. Nucl. Part. Sci., Vol. 24, p. 151, 1974.
76. Poenaru, D. N. and Ivascu, M. S., Particle Emission from Nuclei, Vols. 1-3, CRC Press, Boca Raton, 1988.

77. Wheeler, J. A., "Fission in 1939: The Puzzle and the Promise," Ann. Rev. Nucl. Part. Sci., Vol. 39, p. xiii, 1989.
78. Price, P. B., "Heavy-Particle Radioactivity," Ann. Rev. Nucl. Part. Sci., Vol. 39, p. 19, 1989.
79. Bhandari, B. S., "Resonant Tunneling and the Bimodal Symmetric Fission of ^{258}Fm ," Phys. Rev. Lett., Vol. 66, p. 1034, 25 Feb. 1991.
80. Wagemans, C., The Nuclear Fission Process, CRC Press, Boca Raton, 1991.
81. Harding, A. K., "Physics in Strong Magnetic Fields Near Neutron Stars," Science, Vol. 251, p. 1033, 1 March 1991.
82. Mossberg, T. W., Lewenstein, M. and Gauthier, D. J., "Trapping and Cooling of Atoms in a Vacuum Perturbed in a Frequency-Dependent Manner," Phys. Rev. Lett., Vol. 67, p. 1723, 23 September 1991.
83. Delande, D. and Gay, J. C., "Supersymmetric Factorization for Rydberg Atoms in Parallel Electric and Magnetic Fields," Phys. Rev. Lett., Vol. 66, p. 3237, 24 June 1991.
84. Delande, D., Bommier, A. and Gay, J. C., "Positive-Energy Spectrum of the Hydrogen Atom in a Magnetic Field," Phys. Rev. Lett., Vol. 66, p. 141, 14 January 1991.
85. Iu, C., Welch, G. R., Kash, M. M. and Kleppner, D., "Diamagnetic Rydberg Atom: Confrontation of Calculated and Observed Spectra," Phys. Rev. Lett., Vol. 66, p. 145, 14 January 1991.
86. Pont, M. and Gavrilu, M., "Stabilization of Atomic Hydrogen in Superintense, High-Frequency Laser Fields of Circular Polarization," Phys. Rev. Lett., Vol. 65, p. 2362, 5 November 1990.
87. Greiner, W., Müller, B. and Rafelski, J., Quantum Electrodynamics of Strong Fields, Springer-Verlag, New York, 1985.
88. Greiner, W., editor, Physics of Strong Fields, Plenum Press, New York, 1987.
89. O'Mahony, P. F. and Mota-Furtado, F., "Continuum Spectrum of an Atom or Molecule in a Magnetic Field," Phys. Rev. Lett., Vol. 67, p. 2283, 21 Oct. 1991.
90. Kluger, Y., Eisenberg, J. M., Svetitsky, B., Cooper, F. and Mottola, E., "Pair Production in a Strong Electric Field," Phys. Rev. Lett., Vol. 67, p. 2427, 28 October 1991.
91. Stratton, J. A., Electromagnetic Theory, McGraw-Hill, New York, 1941.
92. Kong, J., Electromagnetic Wave Theory, John Wiley, New York, 1986.
93. Federspiel, F. J., Eisenstein, R. A., Lucas, M. A., MacGibbon, B. E., Mellendorf, K., Nathan, A. M., O'Neill, A. and Wells, D. P., "Proton Compton

Effect: A Measurement of the Electric and Magnetic Polarizabilities of the Proton," Phys. Rev. Lett., Vol. 67, P. 1511, 16 September 1991.

94. Bernard, V. and Kaiser, N., "Chiral Expansion of the Nucleon's Electromagnetic Polarizabilities," Phys. Rev. Lett., Vol. 67, p. 1515, 16 Sept. 1991.

95. Westergaard, H. M., Theory of Elasticity and Plasticity, John Wiley, New York, 1952.

96. Weiss, R. A., Relativistic Thermodynamics, Vols. 1 & 2, Exposition Press, New York, 1976.

97. Emsley, J., The Elements, Oxford Univ. Press, New York, 1991.

98. Palmer, R. and Tollestrup, A. V., "Superconducting Magnet Technology for Accelerators," Ann. Rev. Nucl. Part. Sci., Vol. 34, p. 247, 1984.

99. Date, M. and Kindo, K., "Elementary Excitation in the Haldane State," Phys. Rev. Lett., Vol. 65, p. 1659, 24 September 1990.

100. Katsumata, K., Hori, H., Takeuchi, T., Date, M., Yamagishi, A. and Renard, J. P., Phys. Rev. Lett., Vol. 63, p. 86, 1989.

101. Turchi, P. J., Megagauss Physics and Technology, Plenum Press, New York, 1979.

102. Date, M., editor, High Field Magnetism, North-Holland/Elsevier, New York, 1983.

103. Herlach, F., editor, Strong and Ultrastrong Magnetic Fields and Their Applications, Springer-Verlag, New York, 1985.

104. Fowler, C. M., Cairo, R. S. and Erickson, D. J., Megagauss Technology and Pulsed Power Applications, Plenum Press, New York, 1986.

105. Stephenson, R., Mechanics and Properties of Matter, John Wiley, New York, 1952.

106. Anderson, H., editor, A Physicists Desk Reference, AIP, New York, 1989.

107. Bortignon, P. F., Bracco, A., Brink, D. and Broglia, R. A., "Limiting Temperature for the Existence of Collective Motion in Hot Nuclei," Phys. Rev. Lett., Vol. 67, 9 December 1991.

108. Lerner, R. and Trigg, G., editors, Encyclopedia of Physics, VCH Publishers, New York, 1991.

Table 1. Nuclear Characteristics of the γ Ray Catalyzed Clean Fission of Actinide and Subactinide Elements by Thermal Neutrons

Nucleus	Z^2/A	χ	θ_z^F (degrees)	B^F (10^{12} T)	B_7^F (10^3 T)
^{239}Pu	36.97	1.020	0	0	0
^{233}U	36.33	1.002	0	0	0
^{234}U	36.17	0.998	2.7	3.27	1.34
^{235}U	36.02	0.994	0	0	0
^{231}Pa	35.85	0.989	6.0	7.36	3.02
^{227}Th	35.68	0.984	0	0	0
^{238}U	35.56	0.981	7.8	9.64	3.94
^{228}Th	35.53	0.980	0	0	0
^{232}Th	34.91	0.963	10.9	13.41	5.51
^{208}Pb	32.33	0.892	18.3	22.51	9.41
^{181}Ta	29.44	0.812	23.5	28.91	12.37
^{180}Hf	28.80	0.795	24.4	30.01	12.86
^{152}Sm	25.29	0.698	28.8	35.42	15.61
^{139}La	23.37	0.645	30.8	37.88	16.94
^{127}I	22.12	0.610	32.0	39.36	17.87
^{90}Zr	17.78	0.490	35.5	43.67	21.00
^{88}Sr	16.41	0.453	36.5	44.90	21.67
^{87}Rb	15.74	0.434	37.0	45.51	22.01
^{75}As	14.52	0.401	37.8	46.49	23.05
^{64}Zn	14.06	0.388	38.0	46.74	23.79
^{58}Ni	13.52	0.373	38.4	47.23	24.44
^{63}Cu	13.35	0.369	38.5	47.36	24.17
^{58}Co	12.36	0.341	39.0	47.97	24.75
^{56}Fe	12.07	0.333	39.2	48.22	25.10
^{16}O	4.0	0.110	43.3	53.26	34.16
^{14}N	3.5	0.097	43.5	53.51	35.09
^{12}C	3.0	0.083	43.8	53.87	36.23
^4He	1.0	0.028	44.6	54.86	44.33

Table 2. Nuclear Characteristics of the γ Ray Catalyzed Clean Fission of Actinide and Subactinide Elements by Thermal Neutrons

Nucleus	$10^{10} \zeta_r$	$K_{\frac{87}{92}}^{\frac{87}{92}}$ (10^2 T/deg)	k (10^{20} N/m)	f_r (10^{21} Hz)	λ_r (fm)
^{239}Pu	4.086	5.026	1.744	3.116	96.19
^{233}U	4.104	5.048	1.729	3.143	95.38
^{234}U	4.101	5.044	1.732	3.138	95.51
^{235}U	4.098	5.041	1.734	3.134	95.65
^{231}Pa	4.110	5.055	1.724	3.152	95.10
^{227}Th	4.122	5.070	1.714	3.170	94.55
^{238}U	4.089	5.029	1.741	3.121	96.06
^{228}Th	4.119	5.066	1.717	3.166	94.69
^{232}Th	4.107	5.052	1.727	3.147	95.24
^{208}Pb	4.182	5.144	1.665	3.264	91.84
^{181}Ta	4.280	5.264	1.590	3.419	87.68
^{180}Hf	4.284	5.269	1.587	3.425	87.52
^{152}Sm	4.407	5.421	1.500	3.624	82.72
^{139}La	4.473	5.502	1.456	3.734	80.29
^{127}I	4.541	5.585	1.412	3.848	77.91
^{90}Zr	4.809	5.915	1.259	4.316	69.46
^{88}Sr	4.827	5.937	1.250	4.348	68.94
^{87}Rb	4.836	5.948	1.245	4.365	68.68
^{75}As	4.957	6.097	1.185	4.586	65.37
^{64}Zn	5.090	6.261	1.124	4.835	62.00
^{58}Ni	5.174	6.364	1.088	4.996	60.00
^{63}Cu	5.103	6.277	1.118	4.860	61.68
^{59}Co	5.159	6.346	1.094	4.968	60.34
^{56}Fe	5.205	6.402	1.075	5.055	59.30
^{16}O	6.413	7.888	0.708	7.675	39.06
^{14}N	6.557	8.065	0.677	8.024	37.36
^{12}C	6.728	8.275	0.643	8.448	35.49
^4He	8.080	9.938	0.446	12.183	24.60

Table 3. Nuclear Characteristics of the γ Ray Catalyzed Clean Fission of Actinide and Subactinide Elements by Thermal Neutrons

Nucleus	H_7^F (10^9 amp/m)	E_7^F (10^{12} volts/m)	P_7^F (10^{22} W/m ²)	ϵ_7^F (MeV)	Φ_7^F (10^{34} m ⁻² sec ⁻¹)	n_7^F (10^{23} m ⁻³)
²³⁹ Pu	0	0	0	0	0	0
²³³ U	0	0	0	0	0	0
²³⁴ U	1.07	0.40	0.02	12.98	0.01	0.03
²³⁵ U	0	0	0	0	0	0
²³¹ Pa	2.41	0.91	0.11	13.04	0.05	0.18
²²⁷ Th	0	0	0	0	0	0
²³⁸ U	3.14	1.18	0.19	12.91	0.09	0.30
²²⁸ Th	0	0	0	0	0	0
²³² Th	4.39	1.65	0.36	13.02	0.17	0.58
²⁰⁸ Pb	7.49	2.82	1.06	13.50	0.49	1.63
¹⁶¹ Ta	9.85	3.71	1.83	14.14	0.81	2.70
¹⁸⁰ Hf	10.23	3.86	1.97	14.17	0.87	2.90
¹⁵² Sm	12.42	4.68	2.91	14.99	1.21	4.04
¹³⁹ La	13.48	5.08	3.43	15.44	1.39	4.62
¹²⁷ I	14.23	5.36	3.82	15.92	1.50	4.99
⁹⁰ Zr	16.72	6.30	5.27	17.85	1.84	6.14
⁸⁸ Sr	17.24	6.50	5.61	17.98	1.95	6.49
⁸⁷ Rb	17.52	6.60	5.78	18.05	2.00	6.67
⁷⁵ As	18.34	6.91	6.34	18.97	2.09	6.96
⁶⁴ Zn	18.94	7.14	6.76	20.00	2.11	7.04
⁵⁸ Ni	19.45	7.33	7.13	20.67	2.15	7.18
⁶³ Cu	19.23	7.25	6.97	20.11	2.16	7.22
⁵⁹ Co	19.69	7.42	7.31	20.55	2.22	7.40
⁵⁶ Fe	19.97	7.53	7.51	20.91	2.24	7.48
¹⁶ O	27.18	10.25	13.92	31.75	2.74	9.13
¹⁴ N	27.92	10.53	14.70	33.19	2.76	9.22
¹² C	28.83	10.87	15.67	34.94	2.80	9.34
⁴ He	35.27	13.30	23.45	50.40	2.90	9.69

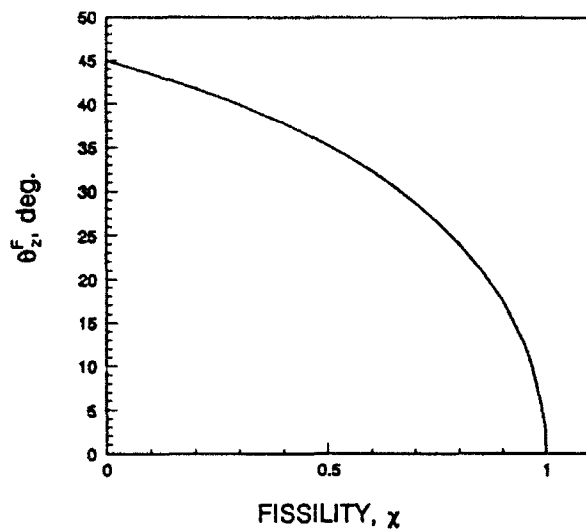


Figure 1

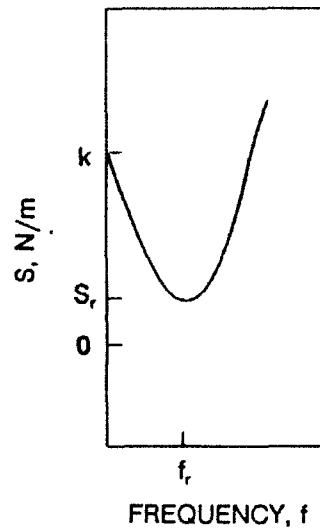


Figure 2

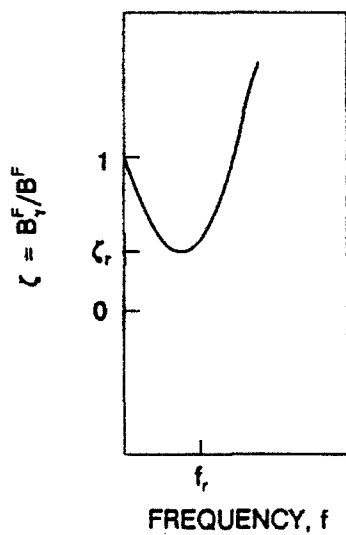


Figure 3

Fig. 1. Internal phase angle of the atomic number that is required for the clean fission of subactinide nuclei versus fissility parameter.

Fig. 2. Frequency dependence of the impedance of an atomic nucleus (not to scale).

Fig. 3. Ratio of dynamic to static magnetic fields required for fission versus frequency (not to scale).

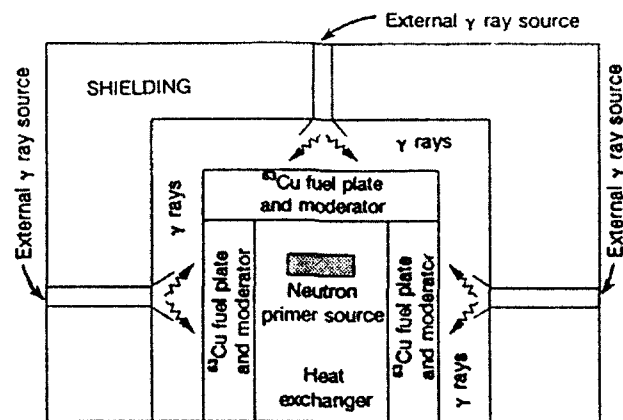


Figure 4. Design concept for a γ ray catalyzed thermal neutron induced clean fission nuclear reactor core.

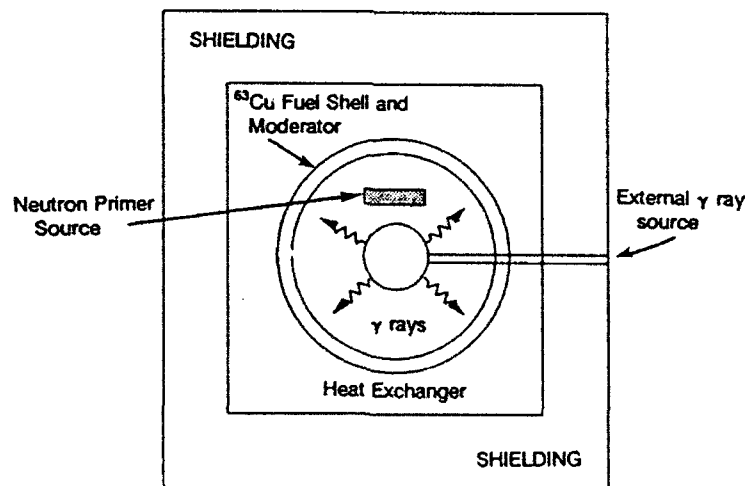


Figure 5. Design concept for γ ray catalyzed thermal neutron clean fission nuclear reactor core.

THERMONUCLEAR REACTIONS IN STRONG GRAVITATIONAL FIELDS

Richard A. Weiss

U.S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. This paper calculates the rates of thermonuclear reactions in gases that are located in gravitational fields or magnetic fields which are sufficiently strong as to produce a partially coherent or totally coherent spacetime state in the gases. Within broken symmetry spacetime the space and time coordinates are represented by complex numbers in an internal space, and therefore the single particle momentum and energy must also be represented as complex numbers in an internal space. The variation of the space and time coordinates and the kinematical variables such as single particle momentum and energy can occur in two limiting ways: a) an incoherent variation in which the magnitudes of these quantities change, and b) a coherent variation wherein these quantities rotate in internal space with fixed magnitudes. These two extremes in the variation of the spacetime coordinates and the kinematical variables determine the limiting forms of the internal energy, pressure and chemical potential of a thermodynamic system because these physical quantities are derived from integrals of distribution functions with respect to complex number single particle momenta or energies. The equations of state for the noninteracting Boltzmann, Fermi and Bose gases are derived for the broken spacetime symmetry state that is induced by a gravitational field. The complex number phase space integrals for the particle number, internal energy and pressure are developed for the partially coherent spacetime state of classical and quantum gases and are then specialized to the limiting case of incoherent and coherent spacetime. The thermonuclear reaction rates for the Boltzmann, Fermi and Bose gases are calculated for the general case of partially coherent spacetime, and then specialized to the cases of coherent and incoherent spacetime states. It is suggested that the measured neutrino deficit for the sun can be explained by the combined effects of the lowered nuclear reaction rates due the effects of the time renormalization group equation and by a reduction in the thermonuclear reaction rates due to a coherent spacetime state which may exist in some regions of the interior of the sun.

1. INTRODUCTION. Element formation in stars is due to a series of thermonuclear reactions.¹⁻⁸ The same reactions are responsible for energy generation in stars whose light is observed and analyzed by scientists on the earth in an attempt to understand nucleosynthesis. The types and rates of nuclear reactions depend on the central temperature and density of stars, and therefore the initial mass of a star ultimately determines the relative abundance of the elements that are synthesized.¹⁻⁸ Only objects with masses greater than $0.08 M_{\odot}$ are stars with thermonuclear reactions in their interiors. Low mass stars with $M < 1 M_{\odot}$ remain main sequence stars that undergo hydrogen burning and convert hydrogen to helium.¹⁻⁸ For stars with $M > 1 M_{\odot}$ helium burning and heavier element burning occurs and the evolutionary outcome depends on the initial stellar mass. For initial masses varying from 1 to $8 M_{\odot}$ the stars can lose enough mass to reach a white dwarf end state with a mass of about $0.6 M_{\odot}$ and an

interior consisting mainly of ^{12}C and degenerate electrons.¹⁻⁸ The maximum mass of a white dwarf is $M \sim 1.4 M_{\odot}$, the Chandrasekhar limit. For stars with initial masses greater than $8 M_{\odot}$ it is not possible to lose sufficient mass during their lifetimes and the cores of these stars undergo a series of contractions to higher temperatures and densities which lead to the thermonuclear production of elements heavier than helium and resulting finally in a core consisting of ^{56}Fe and degenerate electrons.¹⁻⁸ Since ^{56}Fe has the greatest binding energy it cannot undergo subsequent fusion reactions and the mass of the iron core begins to grow. When this core of iron and degenerate electrons exceeds the Chandrasekhar limit a collapse occurs with a rapid electron capture by the protons of the nuclei which leads to neutron rich nuclei, neutron drip and ultimately to a supernova explosion and the possible formation of a neutron star if the remnant stellar nucleus has a mass $M < 3 M_{\odot}$.¹⁻⁸ Supernovae explosions are responsible for the heavy element distribution in the universe.^{4,9-13} The final evolutionary development for supernova remnants with $M > 3 M_{\odot}$ results in perhaps a black hole if classical general relativity is valid at high densities. However, quantum gravity effects may lead to different final collapsed states, but quantum gravity is not yet properly understood. Light elements such as deuterium, lithium and helium were formed mainly during the big bang nucleosynthesis.^{14,15}

Supernova nucleosynthesis in stars results in a distribution of heavy elements throughout the universe in the form of stars and planets. The measured element abundances can be used to develop models of the thermonuclear reactions in stars and these can be used in conjunction with laboratory measured fusion cross sections to understand the temperature, pressure and composition of stellar interiors. There is good agreement between measured abundances of heavy elements and the theoretically predicted abundances based on a standard model of nucleosynthesis.¹⁻¹⁰ Energy production in stars appears ultimately to an observer on the earth as electromagnetic radiation and as particle radiations such as neutrinos. From an analysis of these emanations the astrophysicist attempts to develop physical models of stellar interiors that predict the luminosity and spectrum of the photons and neutrinos emitted from the interior regions. The observed electromagnetic emissions from stellar surfaces agree very well with the theoretical model predictions of the stellar surface temperatures.² However, the agreement between measured and predicted neutrino production is not good. Over the past decade it has become clear that the predicted neutrino flux generated by the standard model of nucleosynthesis is about three times larger than the neutrino flux measured in detectors on the earth.¹⁶⁻²⁴ Several explanations of this discrepancy have been suggested, including oscillation of supposed massive neutrinos.¹⁶⁻²⁹ It is not yet clear whether the discrepancy between the measured and predicted neutrino fluxes is due to errors in the standard nucleosynthesis model or to some other process such as neutrino oscillations during their transit from the sun's interior to the earth.¹⁶⁻²⁹

Several explanations of the experimental neutrino deficit can be given in terms of the internal structure of stars. One idea that has been suggested is that the calculation of the rates of processes that occur in a medium must depend on the state equation of the bulk matter within which the processes occur.^{30,31} Thus the thermonuclear reaction rates within the interior of stars must depend not only on the values of the temperature and density of the stellar gases but also on the state equation of the stellar material.^{30,31} The renormalized relativistic time constant t' of a process such as a thermonuclear reaction occurring in a real gas must be related to the unrenormalized time constant t''

for the process, as given by a standard calculation, by the following gauge invariant time renormalization group equation^{30,31}

$$t' - \beta_E' \partial t' / \partial E' + 3\beta_P' \partial t' / \partial P' = t^{a'} - \beta_E^{a'} \partial t^{a'} / \partial E^{a'} \quad (1A)$$

where the gauge parameters are given by

$$\beta_E' = T/V' (dU'/dT)_{P',V'}, \quad \beta_E^{a'} = T/V' (dU^{a'}/dT)_{P^{a'},V'} \quad (2)$$

$$\beta_P' = d/dV' (P'V')_{U'}, \quad \beta_P^{a'} = d/dV' (P^{a'}V')_{U^{a'}} \quad (3)$$

where T = absolute temperature, V' = volume of space in broken symmetry spacetime, U' and $U^{a'}$ = renormalized and unrenormalized values of the internal energy in broken symmetry spacetime, E' and $E^{a'}$ = renormalized and unrenormalized values of the energy density in broken symmetry spacetime and where P' and $P^{a'}$ = renormalized and unrenormalized values of the pressure in broken symmetry spacetime. The renormalized and unrenormalized (standard) rates of a process are given by $R' = 1/t'$ and $R^{a'} = 1/t^{a'}$ respectively. The time renormalization group equation (1A) is formulated to include the Minkowski metric, and generally predicts that processes occurring within a real medium run slower ($t' > t^{a'}$) than the standard calculations would predict.³⁰ For an ideal gas $\beta_E' = 0$ and $\beta_P' = 0$ so that $t' = t^{a'}$. Quantized versions of the time equation (1A) have been formulated which suggest that energy and pressure fields can have a time structure, and that the rates of processes such as thermonuclear reactions can be depressed or enhanced according to how structures of time can form within the matter in the interior of stars.³¹ The quantized version of the time equation is written as³¹

$$(1 - \omega')(t' - \beta_E' \partial t' / \partial E') + 3\beta_P' \partial t' / \partial P' = 0 \quad (1B)$$

where ω' can be a discrete or continuous eigenvalue. For an ideal gas $\omega' = 1$, $\beta_E' = 0$ and $\beta_P' = 0$.

Probably more than one effect is responsible for the measured neutrino deficit, and equations (1A) and (1B) are only part of the solution. This paper suggests an additional effect that causes another portion of the neutrino deficit, namely that the discrepancy between the experimental and predicted neutrino emission rates from the sun is also due to an overestimation of the neutrino generation rate $R^{a'}$ in the standard stellar models because of the assumption that the stellar interior can be described by an incoherent spacetime Maxwell-Boltzmann gas.² It is suggested in this paper that in fact stellar interiors should be described by a Maxwell-Boltzmann gas in partially coherent or totally coherent spacetime, and that this combined with the time equations (1A) and (1B) if interactions are considered can describe the measured neutrino deficit. For ideal gases the time renormalization group equations (1A) and (1B) give $R' = R^{a'}$, and only the effects of broken symmetry spacetime contribute to the measured neutrino deficit.

A material existing in coherent spacetime is a special case of a material embedded in broken symmetry spacetime having partial coherence. The broken

symmetries of spacetime can be described by writing the spacetime coordinates as complex numbers in an internal space in the following manner³²

$$\bar{t} = t \exp(j\theta_t) \quad (4)$$

$$\bar{x} = x \exp(j\theta_x) \quad \bar{y} = y \exp(j\theta_y) \quad \bar{z} = z \exp(j\theta_z) \quad (5)$$

where θ_t , θ_x , θ_y and θ_z = internal phase angles of time and space. The volume of space is written for the general case of broken symmetry spacetime as³⁰⁻³²

$$V' = \int |d\bar{V}| = \int \sec \beta_{VV} dV = \int \csc \beta_{VV} V d\theta_V \quad (6)$$

where

$$\tan \beta_{VV} = V \partial \theta_V / \partial V \quad (7)$$

For incoherent spacetime³⁰⁻³²

$$\beta_{VV} = 0 \quad V' = V \quad (8)$$

while for coherent spacetime³⁰⁻³²

$$\beta_{VV} = \pi/2 \quad V' = V \theta_V \quad (9)$$

$$\theta_x = \theta_y = \theta_z = \pi/3 \quad \theta_t = \pi/6 \quad (10)$$

Exhaustive treatments of the statistical mechanics of matter for incoherent spacetime can be found in the literature.³³⁻³⁵ This paper extends this literature by treating the statistical mechanics of ideal gases for partial and total coherence of spacetime.

The calculation of the measured pressure and internal energy of a thermodynamic system requires the solution of a renormalization group equation which relates the renormalized relativistic internal energy and pressure to the unrenormalized (standard) values of the internal energy and pressure by the following relativistic trace equation^{30-32,36}

$$\bar{U}' + T(d\bar{U}'/dT)_{\bar{P}', V'} - 3V'd/dV'(\bar{P}'V')_{\bar{U}'} = \bar{U}^a + T(d\bar{U}^a/dT)_{\bar{P}^a, V'} \quad (11A)$$

where \bar{P}' and \bar{U}' = renormalized complex number pressure and internal energy in broken symmetry spacetime, and \bar{P}^a and \bar{U}^a = unrenormalized pressure and internal energy in broken symmetry spacetime. If the complex number renormalized pressure and internal energy are written as³²

$$\bar{P}' = P' \exp(j\theta_P') \quad \bar{U}' = U' \exp(j\theta_U') \quad (12)$$

then the measured pressure and internal energy are given by³²

$$P'_m = P' \cos \theta'_P \quad U'_m = U' \cos \theta'_U \quad (13)$$

For ideal gases (for which $\bar{P}'V' = 2/3\bar{U}'$) it follows from the relativistic trace equation (11A) that³⁶

$$\bar{P}' = \bar{P}^a' \quad \bar{U}' = \bar{U}^a' \quad (14)$$

The quantum version of equation (11A) is written as³¹

$$(1 - \bar{\mu}')[\bar{U}' + T(d\bar{U}'/dT)_{\bar{P}', V'}] - 3V'd/dV'(\bar{P}'V')_{\bar{U}'} = 0 \quad (11B)$$

where $\bar{\mu}'$ = continuous or discrete eigenvalue. For interacting systems the unrenormalized quantities \bar{U}^a' and \bar{P}^a' must first be calculated, and then the relativistic trace equation (11A) must be solved to determine \bar{U}' and \bar{P}' in order to obtain the measured values of the pressure and internal energy given by equation (13). Equations (11A) and (11B) are renormalization group equations that are analogous to the time equations (1A) and (1B) respectively. For a complete analysis equations (1A) and (11A) or equations (1B) and (11B) must be solved jointly to determine the thermonuclear reaction rates in stars whose gases have real state equations. The effects of broken symmetry spacetime appear in all four equations (1A), (1B), (11A) and (11B) through the variable V' . However, in this paper ideal gases are treated and none of these four equations need be considered because $t' = t^a'$ and $\bar{U}' = \bar{U}^a'$ and the superscript "a" will be dropped throughout this paper. Only the effects of broken spacetime symmetry on ideal gas state equations and thermonuclear reaction rates are considered in this paper.

Consider now the velocity and acceleration of a particle located in spacetime with broken internal symmetries. For simplicity only the radial coordinate is considered but similar expressions hold for cartesian coordinates.³² For spacetime with broken internal symmetries the change in the complex number radial coordinate $\bar{\rho}$ and the change in the complex number time \bar{t} are written as³²

$$d\bar{\rho} = \sec \beta_{\rho\rho} d\rho \exp[j(\theta_\rho + \beta_{\rho\rho})] \quad (15)$$

$$= \csc \beta_{\rho\rho} \rho d\theta_\rho \exp[j(\theta_\rho + \beta_{\rho\rho})] \quad (16)$$

$$d\bar{t} = \sec \beta_{tt}^0 dt \exp[j(\theta_t^0 + \beta_{tt}^0)] \quad (17)$$

$$= \csc \beta_{tt}^0 t d\theta_t^0 \exp[j(\theta_t^0 + \beta_{tt}^0)] \quad (18)$$

where

$$\tan \beta_{tt}^0 = t \partial \theta_t^0 / \partial t \quad \tan \beta_{\rho\rho} = \rho \partial \theta_\rho / \partial \rho \quad (19)$$

The velocity can be written as

$$\bar{v} = v \exp(j\theta_v) = d\bar{\rho}/d\bar{t} = d/d\bar{t}[\rho \exp(j\theta_\rho)] \quad (20)$$

where

$$v = \cos \beta_{tt}^{\rho} \sec \beta_{\rho\rho} d\rho/dt \quad (21)$$

$$= \sin \beta_{tt}^{\rho} \csc \beta_{\rho\rho} \rho/t d\theta_{\rho}/d\theta_t^{\rho} \quad (22)$$

$$\theta_v = \theta_{\rho} + \beta_{\rho\rho} - \theta_t^{\rho} - \beta_{tt}^{\rho} \quad (23)$$

For coherent spacetime

$$\beta_{\rho\rho} = \pi/2 \quad \beta_{tt}^{\rho} = \pi/2 \quad (24)$$

so that

$$d\bar{\rho} = j\bar{\rho}d\theta_{\rho} \quad d\bar{t} = j\bar{t}d\theta_t^{\rho} \quad (25)$$

For the coherent spacetime state it follows that

$$v^{\text{coh}} = \rho/t d\theta_{\rho}/d\theta_t^{\rho} \quad \theta_v^{\text{coh}} = \theta_{\rho} - \theta_t^{\rho} \quad (26)$$

where θ_t^{ρ} = internal phase angle of time that is associated with the time variation of the radial coordinate. In general each coordinate x, y, z or r, ϕ, z or ρ, ϕ, ψ has it's own internal phase angle of time $\theta_t^x, \theta_t^y, \theta_t^z$ or $\theta_t^r, \theta_t^{\phi}, \theta_t^z$ or $\theta_t^{\rho}, \theta_t^{\phi}, \theta_t^{\psi}$.

The change in velocity is written as³²

$$d\bar{v} = \sec \beta_{vv} dv \exp[j(\theta_v + \beta_{vv})] \quad (27)$$

$$= \csc \beta_{vv} v d\theta_v \exp[j(\theta_v + \beta_{vv})] \quad (28)$$

where

$$\tan \beta_{vv} = v \partial \theta_v / \partial v \quad (29)$$

and therefore the acceleration is given by

$$\bar{a} = a \exp(j\theta_a) = d\bar{v}/d\bar{t} \quad (30)$$

where

$$a = \cos \beta_{tt}^\rho \sec \beta_{vv} dv/dt \quad (31)$$

$$= \cos \beta_{tt}^\rho \sec \beta_{vv} d/dt (\cos \beta_{tt}^\rho \sec \beta_{\rho\rho} d\rho/dt) \quad (32)$$

$$= \sin \beta_{tt}^\rho \csc \beta_{vv} v/t d\theta_v/d\theta_t^\rho \quad (33)$$

$$= \sin^2 \beta_{tt}^\rho \csc \beta_{vv} \csc \beta_{\rho\rho} \rho/t^2 d\theta_v/d\theta_t^\rho d\theta_\rho/d\theta_t^\rho \quad (34)$$

$$\theta_a = \theta_v + \beta_{vv} - \theta_t^\rho - \beta_{tt}^\rho \quad (35)$$

$$= \theta_\rho + \beta_{vv} + \beta_{\rho\rho} - 2(\theta_t^\rho + \beta_{tt}^\rho) \quad (36)$$

Equation (34) can also be written as

$$a = \sin^2 \beta_{tt}^\rho \csc \beta_{\rho\rho} \csc \beta_{vv} \rho/t^2 [d\theta_\rho/d\theta_t^\rho - 1 + d/d\theta_t^\rho (\beta_{\rho\rho} - \beta_{tt}^\rho)] d\theta_\rho/d\theta_t^\rho \quad (37)$$

For coherent spacetime equation (24) is valid and equation (29) becomes

$$\tan \beta_{vv}^{\text{coh}} = d\theta_\rho/d\theta_t^\rho (d\theta_\rho/d\theta_t^\rho - 1) / (d^2\theta_\rho/d\theta_t^{\rho 2}) \quad (38)$$

while from equations (24) and (35) through (38) it follows that

$$a^{\text{coh}} = \rho/t^2 [(d\theta_\rho/d\theta_t^\rho)^2 (d\theta_\rho/d\theta_t^\rho - 1)^2 + (d^2\theta_\rho/d\theta_t^{\rho 2})^2]^{1/2} \quad (39)$$

$$\theta_a^{\text{coh}} = \theta_\rho - 2\theta_t^\rho + \beta_{vv}^{\text{coh}} - \pi/2 \quad (40)$$

These equations simplify further if $\theta_\rho = b\theta_t^\rho + e$ with $b \leq 1$

$$\beta_{vv}^{\text{coh}} = -\pi/2 \quad d^2\theta_\rho/d\theta_t^{\rho 2} = 0 \quad d\theta_\rho/d\theta_t^\rho = b \quad (41)$$

and equation (28) becomes for coherent spacetime

$$d\bar{v} = \bar{v} d\theta_v \quad (42)$$

while equations (39) and (40) become

$$a^{\text{coh}} = \rho/t^2 |b(b-1)| = \rho/t^2 b(1-b) \quad (43)$$

$$\theta_a^{\text{coh}} = \theta_\rho - 2\theta_t^\rho - \pi \quad (44)$$

because in general $b \leq 1$.

The single particle momentum is written as

$$\bar{p} = p \exp(j\theta_p) = m\bar{v} \quad (45)$$

$$p = mv \quad \theta_p = \theta_v \quad (46)$$

so that equations (27) and (28) give

$$d\bar{p} = \sec \beta_{pp} dp \exp[j(\theta_p + \beta_{pp})] \quad (47)$$

$$= \csc \beta_{pp} p d\theta_p \exp[j(\theta_p + \beta_{pp})] \quad (48)$$

where

$$\tan \beta_{pp} = p \partial \theta_p / \partial p \quad (49)$$

The single particle energy is written as

$$\bar{\epsilon} = \epsilon \exp(j\theta_\epsilon) = \bar{p}^2 / (2m) \quad (50)$$

so that

$$\epsilon = p^2 / (2m) \quad \theta_\epsilon = 2\theta_p = 2\theta_v \quad (51)$$

where θ_v is given by equation (23). The variation of the complex number single particle energy is then written as

$$d\bar{\epsilon} = \sec \beta_{\epsilon\epsilon} d\epsilon \exp[j(\theta_\epsilon + \beta_{\epsilon\epsilon})] \quad (52)$$

$$= \csc \beta_{\epsilon\epsilon} \epsilon d\theta_\epsilon \exp[j(\theta_\epsilon + \beta_{\epsilon\epsilon})] \quad (53)$$

where

$$\tan \beta_{\epsilon\epsilon} = \epsilon \partial \theta_\epsilon / \partial \epsilon \quad (54)$$

For a coherent spacetime state described by equations (24) and (41) it follows that

$$\beta_{pp} = \pi/2 \quad \beta_{\epsilon\epsilon} = \pi/2 \quad (55)$$

and

$$d\bar{p} = j\bar{p}d\theta_p \quad d\bar{\epsilon} = j\bar{\epsilon}d\theta_\epsilon \quad (56)$$

where the magnitudes of the single particle momentum and energy are constants, $p = p_c$ and $\epsilon = \epsilon_c$. Equation (56) will be used in Sections 2 through 5 to evaluate the basic complex number integrals of the statistical mechanics of Boltzmann, Fermi and Bose gases for the case of coherent spacetime.

The state equations and thermonuclear reaction rates associated with the

ideal Boltzmann, Fermi and Bose gases in incoherent spacetime are given by well known standard formalisms.¹⁻³ This paper extends these calculations to the cases of partially coherent and totally coherent spacetime that are associated with matter in strong gravitational fields. Specifically, Section 2 considers the determination of the chemical potential, internal energy and pressure of an ideal Boltzmann gas that is located in spacetime with broken internal symmetries, Section 3 studies the state equation of a noninteracting Fermi gas in a spacetime with partial or total coherence, Section 4 treats the statistical mechanics of an ideal Bose gas in broken symmetry spacetime, and finally Section 5 evaluates the thermonuclear reaction rates that are expected to occur in ideal Boltzmann, Fermi and Bose gases that are located in the partially coherent or totally coherent spacetime that exists in the presence of strong gravitational fields.

2. STATISTICAL MECHANICS OF A BOLTZMANN GAS IN A STRONG GRAVITATIONAL FIELD. This section examines the statistical thermodynamics of a noninteracting Boltzmann gas that is located in a gravitational field or other external field, such as an electromagnetic field, that induces a broken symmetry in the local space and time coordinates. The Boltzmann gas is important for the study of thermonuclear reactions in ordinary stars where the nuclei can be described by an ideal classical gas. The statistical mechanics of an ideal nonrelativistic Boltzmann gas that is located in incoherent spacetime is exhaustively treated in the literature.³³⁻³⁵ This section calculates the chemical potential, internal energy and pressure of an ideal Boltzmann gas for the case of a partially coherent spacetime and for a totally coherent spacetime.

A. Particle Number and Chemical Potential.

It is assumed that the number of particles \bar{n} is a complex number in internal space, like the complex magnetic quantum number \bar{M}' of Reference 32, because it represents a quantum number for a wave function that is expressed in terms of a complex number coordinate $\bar{\zeta}$ in internal space as follows

$$\bar{\psi} = \bar{A} \exp(j\bar{n}\bar{\zeta}) \quad \bar{n} = n \exp(j\theta_n) \quad \bar{\zeta} = \zeta \exp(j\theta_\zeta) \quad (57)$$

where $\bar{\psi}$, \bar{n} and $\bar{\zeta}$ = complex number wave function, particle number and internal space coordinate respectively. Then by the same argument given in Reference 32 for the complex magnetic quantum number \bar{M} it follows that the requirement for periodicity of the wave function gives

$$\bar{n}\bar{\zeta} = n\zeta \quad \theta_n = -\theta_\zeta \quad (58)$$

The measured internal space coordinate is given by³²

$$\zeta_m = \zeta \cos \theta_\zeta \quad (59)$$

The requirement that the wave function be periodic in terms of the measured internal space coordinate gives the following condition³²

$$n\zeta = (n/\cos \theta_\zeta)\zeta_m = N\zeta_m \quad (60)$$

where $N = \text{integer}$, so that the complex particle number can be written

$$\bar{n} = n \exp(j\theta_n) \quad (61)$$

$$= N \cos \theta_n \exp(j\theta_n) = N \cos \theta_\zeta \exp(-j\theta_\zeta) \quad (62)$$

where $N = \text{integer number of particles}$. Therefore

$$n = N \cos \theta_n = N \cos \theta_\zeta \quad (63)$$

and the real and imaginary parts of the particle number are written as

$$n_R = N \cos^2 \theta_n \quad (64)$$

$$n_I = N \cos \theta_n \sin \theta_n \quad (65)$$

The measured particle number is then given by

$$n_m = N \cos^2 \theta_n = n \cos \theta_n \quad (66)$$

The quantities N and θ_n are assumed to be known, and equations (61) through (66) immediately determine \bar{n} , n , n_R , n_I and n_m . These equations are formally identical to the equations for the complex magnetic quantum number \bar{M} and the integer magnetic quantum number m with the correspondence $\bar{n} \leftrightarrow \bar{M}$ and $N \leftrightarrow m$ if N is allowed to have positive and negative integer values, or if N is allowed to have only positive integer values these equations are homologous to the equations for the complex magnetic quantum number \bar{M}' and the integer quantum number $|m|$ with the correspondence $\bar{n} \leftrightarrow \bar{M}'$ and $N \leftrightarrow |m|$.³²

The total particle number can also be written as an integral of the Boltzmann distribution function over the space of single particle momenta and energies.³³⁻³⁵ The generalization to the case of complex number momenta and energies is written

$$\bar{n} = \int_F \bar{z} \exp(-\beta \bar{\epsilon}_p) \quad (67)$$

$$= 4\pi V' \bar{z}/h^3 \int \bar{p}^2 \exp[-\beta \bar{p}^2/(2m)] d\bar{p} \quad (68)$$

$$= V' 2\pi(2m)^{3/2} \bar{z}/h^3 \int \bar{\epsilon}^{1/2} \exp(-\beta \bar{\epsilon}) d\bar{\epsilon} \quad (69)$$

where $V' = \text{broken symmetry volume given by equation (6)}$, $\bar{z} = \text{complex number fugacity}$, $h = \text{Planck's constant}$, $\bar{p} = \text{complex number momentum integration variable}$, $\beta = 1/(kT)$, $T = \text{absolute temperature}$, $m = \text{mass of particles}$ and $\bar{\epsilon}_p = \bar{\epsilon} = \bar{p}^2/(2m) = \text{complex number kinetic energy per particle which in equation (69) appears as an integration variable}$. The complex number fugacity is related to the complex number chemical potential by the following generalization of the standard result³³⁻³⁵

$$\bar{z} = \exp(\beta \bar{\mu}) \quad \bar{\mu} = \mu \exp(j\theta_\mu) \quad (70)$$

where $\bar{\mu}$ = complex number chemical potential. From equation (70) it follows that

$$z = \exp(\beta \mu \cos \theta_\mu) \quad \ln z = \beta \mu \cos \theta_\mu \quad (71)$$

$$\theta_z = \beta \mu \sin \theta_\mu \quad (72)$$

Inverting equations (71) and (72) gives

$$\tan \theta_\mu = \theta_z (\ln z)^{-1} \quad (73)$$

$$\mu = \beta^{-1} (\theta_z^2 + \ln^2 z)^{1/2} \quad (74)$$

which gives the chemical potential in terms of the fugacity.

Equation (69) can be written as

$$\bar{n} = V' f \bar{z} \bar{A} = V' f z A \exp[j(\theta_z + \theta_A)] \quad (75)$$

$$n = V' f z A \quad \theta_n = \theta_z + \theta_A \quad (76)$$

where

$$f = 2\pi (2m)^{3/2} / h^3 \quad (77)$$

and where

$$\bar{A} = A_R + jA_I = A \exp(j\theta_A) = \int \bar{\epsilon}^{1/2} \exp(-\beta \bar{\epsilon}) d\bar{\epsilon} \quad (78)$$

$$A = (A_R^2 + A_I^2)^{1/2} \quad \tan \theta_A = A_I / A_R \quad (79)$$

The single particle kinetic energy integration variable is written as

$$\bar{\epsilon} = \epsilon \exp(j\theta_\epsilon) \quad (80)$$

which gives

$$d\bar{\epsilon} = \sec \beta_{\epsilon\epsilon} \exp[j(\theta_\epsilon + \beta_{\epsilon\epsilon})] d\epsilon \quad (81)$$

$$= \epsilon \csc \beta_{\epsilon\epsilon} \exp[j(\theta_\epsilon + \beta_{\epsilon\epsilon})] d\theta_\epsilon \quad (82)$$

where

$$\tan \beta_{\epsilon\epsilon} = \epsilon \partial \theta_\epsilon / \partial \epsilon \quad (83)$$

Then equation (78) can be written as

$$\bar{A} = \int_0^{\infty} \epsilon^{1/2} \sec \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \exp[j(3/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon})] d\epsilon \quad (84)$$

$$= \int_0^{\pi/3} \epsilon^{3/2} \csc \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \exp[j(3/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon})] d\theta_{\epsilon} \quad (85)$$

so that the component integrals become

$$A_R = \int_0^{\infty} \epsilon^{1/2} \sec \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \cos(3/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\epsilon \quad (86)$$

$$= \int_0^{\pi/3} \epsilon^{3/2} \csc \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \cos(3/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\theta_{\epsilon} \quad (87)$$

$$A_I = \int_0^{\infty} \epsilon^{1/2} \sec \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \sin(3/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\epsilon \quad (88)$$

$$= \int_0^{\pi/3} \epsilon^{3/2} \csc \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \sin(3/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\theta_{\epsilon} \quad (89)$$

The values of A and θ_A are then obtained from equation (79). Equations (63), (76) and (79) can be used to determine z and θ_z in the following manner

$$z = n/(V'fA) \quad (90)$$

$$= (N \cos \theta_n)/(V'fA)$$

$$= (n_m/\cos \theta_n)/(V'fA)$$

$$\theta_z = \theta_n - \theta_A \quad (91)$$

where f is defined in equation (77), and where \bar{n} and n are related to the integer particle number N by equations (61) through (63). Finally, equations (73) and (74) can be used to determine θ_n and μ respectively. The measured particle number is given by equation (66). Note that for $\theta_n = 0$ equation (76) shows that the fugacity can still have an internal phase angle.

All internal phase angles are set equal to zero for the case of incoherent spacetime. For this case equation (75) becomes

$$N = Vfz_{inc} A^{inc} \quad (92)$$

where from equation (78)

$$A^{inc} = \int_0^{\infty} \epsilon^{1/2} \exp(-\beta\epsilon) d\epsilon = 1/2\sqrt{\pi}\beta^{-3/2} = 1/2\sqrt{\pi}(kT)^{3/2} \quad (93)$$

Then equations (92) and (93) can be used to determine z_{inc} . These results are the standard equations for the ideal Boltzmann gas.³³⁻³⁵

Consider now the case of incoherent spacetime when the angle of the single particle energy is given by $\theta_\epsilon = \theta_\epsilon^i = \text{constant}$, then equations (86) and (88) are written as

$$\begin{aligned} A_R^{\text{inc}} &= \int_0^\infty \epsilon^{1/2} \exp(-\beta\epsilon \cos \theta_\epsilon^i) \cos(3/2\theta_\epsilon^i - \beta\epsilon \sin \theta_\epsilon^i) d\epsilon \\ &= I_1 \cos(3/2\theta_\epsilon^i) + I_2 \sin(3/2\theta_\epsilon^i) \end{aligned} \quad (94)$$

$$\begin{aligned} A_I^{\text{inc}} &= \int_0^\infty \epsilon^{1/2} \exp(-\beta\epsilon \cos \theta_\epsilon^i) \sin(3/2\theta_\epsilon^i - \beta\epsilon \sin \theta_\epsilon^i) d\epsilon \\ &= I_1 \sin(3/2\theta_\epsilon^i) - I_2 \cos(3/2\theta_\epsilon^i) \end{aligned} \quad (95)$$

where

$$I_1 = \int_0^\infty \epsilon^{1/2} \exp(-\beta\epsilon \cos \theta_\epsilon^i) \cos(\beta\epsilon \sin \theta_\epsilon^i) d\epsilon \quad (96)$$

$$I_2 = \int_0^\infty \epsilon^{1/2} \exp(-\beta\epsilon \cos \theta_\epsilon^i) \sin(\beta\epsilon \sin \theta_\epsilon^i) d\epsilon \quad (97)$$

Setting $\epsilon = x^2$ allows these integrals to be written as

$$I_1 = 2 \int_0^\infty x^2 \exp(-\beta x^2 \cos \theta_\epsilon^i) \cos(\beta x^2 \sin \theta_\epsilon^i) dx \quad (98)$$

$$I_2 = 2 \int_0^\infty x^2 \exp(-\beta x^2 \cos \theta_\epsilon^i) \sin(\beta x^2 \sin \theta_\epsilon^i) dx \quad (99)$$

These integrals are evaluated in tables of integrals with the result³⁷

$$I_1 = 1/2\sqrt{\pi}\beta^{-3/2} \cos(3/2\theta_\epsilon^i) \quad (100)$$

$$I_2 = 1/2\sqrt{\pi}\beta^{-3/2} \sin(3/2\theta_\epsilon^i) \quad (101)$$

Combining equations (94), (95), (100) and (101) gives

$$A_R^{\text{inc}} = 1/2\sqrt{\pi}\beta^{-3/2} \quad (102)$$

$$A_I^{\text{inc}} = 0 \quad (103)$$

so that within the approximation $\theta_\epsilon = \theta_\epsilon^i = \text{constant}$ the internal phase angle θ_ϵ^i has no effect on the values of the integrals which are now just equivalent to equation (93) for the case when all internal phase angles are set equal to zero.

Consider the representation of the particle number and the calculation of the chemical potential for a Boltzmann gas in a coherent spacetime state. For a coherent spacetime state $\beta_{\epsilon\epsilon} = \pi/2$ and equation (85) becomes

$$\bar{A}^{\text{coh}} = A^{\text{coh}} \exp(j\theta_A^{\text{coh}}) \quad (104)$$

$$= \epsilon_c^{3/2} \bar{W} = \epsilon_c^{3/2} W \exp(j\theta_W) = \epsilon_c^{3/2} (W_R + jW_I)$$

$$A^{\text{coh}} = \epsilon_c^{3/2} W \quad \theta_A^{\text{coh}} = \theta_W \quad (105)$$

where $\epsilon = \epsilon_c$ = constant magnitude of the single particle kinetic energy, and where

$$\bar{W} = j \int_0^{\pi/3} \exp(j3/2\theta_\epsilon) \exp(-\beta\bar{\epsilon}) d\theta_\epsilon \quad (106)$$

$$= j \int_0^{\pi/3} \exp(-\beta\epsilon_c \cos \theta_\epsilon) \exp[j(3/2\theta_\epsilon - \beta\epsilon_c \sin \theta_\epsilon)] d\theta_\epsilon$$

$$W_R = \int_0^{\pi/3} \exp(-\beta\epsilon_c \cos \theta_\epsilon) \sin(\beta\epsilon_c \sin \theta_\epsilon - 3/2\theta_\epsilon) d\theta_\epsilon \quad (107)$$

$$W_I = \int_0^{\pi/3} \exp(-\beta\epsilon_c \cos \theta_\epsilon) \cos(\beta\epsilon_c \sin \theta_\epsilon - 3/2\theta_\epsilon) d\theta_\epsilon \quad (108)$$

where the j that appears in equation (106) results from $\beta_{\epsilon\epsilon} = \pi/2$ in equation (85) and agrees with the results in equation (56). The particle number for the coherent spacetime state is then obtained from equations (75) and (104) to be

$$\bar{n} = V\theta_V f \bar{z}_{\text{coh}} \bar{A}^{\text{coh}} = n \exp(j\theta_n) \quad (109)$$

$$= V\theta_V f \epsilon_c^{3/2} \bar{z}_{\text{coh}} \bar{W}$$

$$= V\theta_V f \epsilon_c^{3/2} z_{\text{coh}} W \exp[j(\theta_z^{\text{coh}} + \theta_W)]$$

where $\bar{z}_{\text{coh}} = z_{\text{coh}} \exp(j\theta_z^{\text{coh}})$ and

$$W = (W_R^2 + W_I^2)^{1/2} \quad \tan \theta_W = W_I/W_R \quad (110)$$

Then

$$n = V\theta_V f \epsilon_c^{3/2} z_{\text{coh}} W \quad \theta_n = \theta_z^{\text{coh}} + \theta_W \quad (111)$$

From equations (63) and (111) it follows that the components of the complex number fugacity are given by

$$z_{\text{coh}} = (N \cos \theta_n) / (V\theta_V f A^{\text{coh}}) \quad (112)$$

$$= (N \cos \theta_n) / (V\theta_V f \epsilon_c^{3/2} W)$$

$$\theta_z^{\text{coh}} = \theta_n - \theta_W \quad (113)$$

Then equations (73) and (74) can be used to determine the magnitude and the internal phase angle of the chemical potential in coherent spacetime μ_{coh} and $\theta_{\mu}^{\text{coh}}$ which gives $\bar{\mu}_{\text{coh}} = \mu_{\text{coh}} \exp(j\theta_{\mu}^{\text{coh}})$. The upper integration limit of $\pi/3$ in equations (106) through (108) is obtained from equations (10), (26) and (51) to be $\theta_{\epsilon} = 2\theta_V = 2(\theta_{\rho} - \theta_t) = 2(\pi/3 - \pi/6) = \pi/3$.

B. Internal Energy of Boltzmann Gas.

In the presence of an external field which breaks the spacetime coordinate symmetry of the vacuum, the internal energy of a Boltzmann gas can be written as the following complex number generalization of the standard scalar result³³⁻³⁵

$$\begin{aligned} \bar{U} &= U \exp(j\theta_U) = 4\pi V' \bar{z} / h^3 \int \bar{p}^2 \bar{p}^2 / (2m) \exp[-\beta \bar{p}^2 / (2m)] d\bar{p} \\ &= V' f \bar{z} \bar{B} \end{aligned} \quad (114)$$

where

$$\bar{B} = B_R + jB_I = B \exp(j\theta_B) = \int \bar{\epsilon}^{3/2} \exp(-\beta \bar{\epsilon}) d\bar{\epsilon} \quad (115)$$

where the broken symmetry volume V' is given by equation (6), f is given by equation (77), and where \bar{z} is given by equation (70). For partial coherence the integral in equation (115) can be written as

$$\bar{B} = \int_0^{\infty} \epsilon^{3/2} \sec \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \exp[j(5/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon})] d\epsilon \quad (116)$$

$$= \int_0^{\pi/3} \epsilon^{5/2} \csc \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \exp[j(5/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon})] d\theta_{\epsilon} \quad (117)$$

and therefore

$$B_R = \int_0^{\infty} \epsilon^{3/2} \sec \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \cos(5/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\epsilon \quad (118)$$

$$= \int_0^{\pi/3} \epsilon^{5/2} \csc \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \cos(5/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\theta_{\epsilon} \quad (119)$$

$$B_I = \int_0^{\infty} \epsilon^{3/2} \sec \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \sin(5/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\epsilon \quad (120)$$

$$= \int_0^{\pi/3} \epsilon^{5/2} \csc \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \sin(5/2\theta_{\epsilon} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\theta_{\epsilon} \quad (121)$$

Equation (114) can be written as

$$U = V' f z B \quad \theta_U = \theta_z + \theta_B \quad (122)$$

where

$$B = (B_R^2 + B_I^2)^{1/2} \quad \tan \theta_B = B_I/B_R \quad (123)$$

The average kinetic energy per particle is then given by equations (75) and (114) to be

$$\bar{\epsilon}_{av} = \bar{U}/\bar{n} = \bar{B}/\bar{A} = B/A \exp[j(\theta_B - \theta_A)] \quad (124)$$

$$\epsilon_{av} = U/n = B/A \quad \theta_{\epsilon}^{av} = \theta_B - \theta_A \quad (125)$$

Then the real and imaginary parts of $\bar{\epsilon}_{av}$ are given by

$$\epsilon_{avR} = \epsilon_{av} \cos \theta_{\epsilon}^{av} = (A_R B_R + A_I B_I)/(A_R^2 + A_I^2) \quad (126)$$

$$\epsilon_{avI} = \epsilon_{av} \sin \theta_{\epsilon}^{av} = (A_R B_I - A_I B_R)/(A_R^2 + A_I^2) \quad (127)$$

where A_R and A_I are given by equations (86) through (89).

For the case of incoherent spacetime all internal phase angles are set equal to zero and equation (122) gives

$$U^{inc} = V f z_{inc} B^{inc} \quad (128)$$

where f is given by equation (77), and where equation (115) becomes

$$\begin{aligned} B^{inc} &= \int_0^{\infty} \epsilon^{3/2} \exp(-\beta \epsilon) d\epsilon \\ &= \beta^{-5/2} \Gamma(5/2) = 3/4 \sqrt{\pi} \beta^{-5/2} = 3/4 \sqrt{\pi} (kT)^{5/2} \end{aligned} \quad (129)$$

which is the standard result.³³⁻³⁵

If the internal phase angle of the single particle energy is assumed to have a constant value $\theta_{\epsilon} = \theta_{\epsilon}^1 = \text{constant}$, the integrals in equations (118) and (120) become

$$B_R^{inc} = \int_0^{\infty} \epsilon^{3/2} \exp(-\beta \epsilon \cos \theta_{\epsilon}^1) \cos(5/2 \theta_{\epsilon}^1 - \beta \epsilon \sin \theta_{\epsilon}^1) d\epsilon \quad (130)$$

$$= I_3 \cos(5/2 \theta_{\epsilon}^1) + I_4 \sin(5/2 \theta_{\epsilon}^1) \quad (131)$$

$$B_I^{inc} = \int_0^{\infty} \epsilon^{3/2} \exp(-\beta \epsilon \cos \theta_{\epsilon}^1) \sin(5/2 \theta_{\epsilon}^1 - \beta \epsilon \sin \theta_{\epsilon}^1) d\epsilon \quad (132)$$

$$= I_3 \sin(5/2 \theta_{\epsilon}^1) - I_4 \cos(5/2 \theta_{\epsilon}^1) \quad (133)$$

where

$$I_3 = \int_0^{\infty} \epsilon^{3/2} \exp(-\beta \epsilon \cos \theta_{\epsilon}^i) \cos(\beta \epsilon \sin \theta_{\epsilon}^i) d\epsilon \quad (134)$$

$$I_4 = \int_0^{\infty} \epsilon^{3/2} \exp(-\beta \epsilon \cos \theta_{\epsilon}^i) \sin(\beta \epsilon \sin \theta_{\epsilon}^i) d\epsilon \quad (135)$$

By integration by parts it is easy to show that

$$I_3 = 3/2\beta^{-1} (I_1 \cos \theta_{\epsilon}^i - I_2 \sin \theta_{\epsilon}^i) \quad (136)$$

$$= 3/4\sqrt{\pi}\beta^{-5/2} \cos(5/2\theta_{\epsilon}^i) \quad (137)$$

$$I_4 = 3/2\beta^{-1} (I_1 \sin \theta_{\epsilon}^i + I_2 \cos \theta_{\epsilon}^i) \quad (138)$$

$$= 3/4\sqrt{\pi}\beta^{-5/2} \sin(5/2\theta_{\epsilon}^i) \quad (139)$$

where I_1 and I_2 are given by equations (100) and (101). Combining equations (131), (133), (137) and (139) gives

$$B_R^{inc} = 3/4\sqrt{\pi}\beta^{-5/2} \quad (140)$$

$$B_I^{inc} = 0 \quad (141)$$

and it is seen that within the approximation $\theta_{\epsilon} = \theta_{\epsilon}^i = \text{constant}$ the phase angle θ_{ϵ}^i does not enter the final expression for the internal energy of a Boltzmann gas, and the result in equation (140) is identical to that in equation (129) for the case when all internal phase angles are set equal to zero.

For the case of coherent spacetime $\beta_{\epsilon\epsilon} = \pi/2$ and $V' = V\theta_V$ and equations (114) and (117) become

$$\bar{U}^{coh} = V\theta_V f \bar{z}_{coh} \bar{B}^{coh} \quad (142)$$

$$= V\theta_V f \epsilon_c^{5/2} \bar{z}_{coh} \bar{Q}$$

$$= V\theta_V f \epsilon_c^{5/2} z_{coh} Q \exp[j(\theta_z^{coh} + \theta_Q)]$$

$$U^{coh} = V\theta_V f \epsilon_c^{5/2} z_{coh} Q \quad (143)$$

$$\theta_U^{coh} = \theta_z^{coh} + \theta_Q \quad (144)$$

where $\epsilon_c = \text{constant value of the magnitude of the single particle kinetic energy, and where}$

$$\bar{B}^{\text{coh}} = \epsilon_c^{5/2} \bar{Q} = \epsilon_c^{5/2} (Q_R + jQ_I) = \epsilon_c^{5/2} Q \exp(j\theta_Q) \quad (145)$$

$$Q = (Q_R^2 + Q_I^2)^{1/2} \quad \tan \theta_Q = Q_I/Q_R \quad (146)$$

with

$$\bar{Q} = j \int_0^{\pi/3} \exp(-\beta \epsilon_c \cos \theta_\epsilon) \exp[j(5/2\theta_\epsilon - \beta \epsilon_c \sin \theta_\epsilon)] d\theta_\epsilon \quad (147)$$

$$Q_R = \int_0^{\pi/3} \exp(-\beta \epsilon_c \cos \theta_\epsilon) \sin(\beta \epsilon_c \sin \theta_\epsilon - 5/2\theta_\epsilon) d\theta_\epsilon \quad (148)$$

$$Q_I = \int_0^{\pi/3} \exp(-\beta \epsilon_c \cos \theta_\epsilon) \cos(\beta \epsilon_c \sin \theta_\epsilon - 5/2\theta_\epsilon) d\theta_\epsilon \quad (149)$$

The average energy per particle for the Boltzmann gas in coherent spacetime is obtained from equation (109) and (142) to be

$$\bar{\epsilon}_{\text{av}}^{\text{coh}} = \bar{U}^{\text{coh}}/\bar{n}^{\text{coh}} = \bar{B}^{\text{coh}}/\bar{A}^{\text{coh}} = \epsilon_c \bar{Q}/\bar{W} \quad (150)$$

$$\epsilon_{\text{av}}^{\text{coh}} = \epsilon_c Q/W \quad \theta_\epsilon^{\text{coh}} = \theta_Q - \theta_W \quad (151)$$

and the real and imaginary parts of the average internal energy per particle are given by

$$\epsilon_{\text{avR}}^{\text{coh}} = \epsilon_c Q/W \cos(\theta_Q - \theta_W) \quad (152)$$

$$= \epsilon_c (W_R Q_R + W_I Q_I) / (W_R^2 + W_I^2)$$

$$\epsilon_{\text{avI}}^{\text{coh}} = \epsilon_c Q/W \sin(\theta_Q - \theta_W) \quad (153)$$

$$= \epsilon_c (W_R Q_I - W_I Q_R) / (W_R^2 + W_I^2)$$

where \bar{W} is given by equation (106), W_R and W_I are given by equations (107) and (108), and where W and θ_W are given in equation (110).

C. Pressure of a Broken Symmetry Boltzmann Gas.

The pressure of a Boltzmann gas that is located in a broken symmetry spacetime, such as may be induced by a strong gravitational field, can be calculated from the following generalization of a standard result³³⁻³⁵

$$\bar{n} = V' \bar{z} \partial / \partial \bar{z} [\bar{P}/(kT)] = V' \bar{z} \partial / \partial \bar{z} (\beta \bar{P}) \quad (154)$$

Combining equations (75) and (154) gives immediately

$$\bar{P} = \bar{n}/V' kT \quad (155)$$

where V' is given by equation (6). Combining equations (61) and (155) gives

$$\bar{P} = N/V'kT \cos \theta_n \exp(j\theta_n) \quad (156)$$

If β_{VV} is constant it follows from equation (6) that

$$V' = V \sec \beta_{VV} = V\theta_V \csc \beta_{VV} \quad (157)$$

so that for this special case

$$\bar{P} = N/VkT \cos \beta_{VV} \cos \theta_n \exp(j\theta_n) \quad (158)$$

$$= N/(V\theta_V)kT \sin \beta_{VV} \cos \theta_n \exp(j\theta_n) \quad (159)$$

where N = integer number of particles in a container. The factors $\cos \beta_{VV}$ or $\sin \beta_{VV}$ arise from the broken symmetry of spacetime.³² The factor $\cos \theta_n$ arises from the assumed periodicity of the quantum mechanical wave function in the measured internal space coordinates. The real and imaginary parts of the complex number pressure are given by

$$P_R = N/VkT \cos \beta_{VV} \cos^2 \theta_n \quad (160)$$

$$= N/(V\theta_V)kT \sin \beta_{VV} \cos^2 \theta_n \quad (161)$$

$$P_I = N/VkT \cos \beta_{VV} \cos \theta_n \sin \theta_n \quad (162)$$

$$= N/(V\theta_V)kT \sin \beta_{VV} \cos \theta_n \sin \theta_n \quad (163)$$

The measured pressure is given by $P_m = P_R$.

The incoherent spacetime pressure of a Boltzmann gas is obtained from equation (160) with $\beta_{VV} = 0$ and $\theta_n = 0$ which gives the standard result³³⁻³⁵

$$P_m = N/VkT \quad (164)$$

The coherent spacetime form of the pressure of an ideal Boltzmann gas is obtained from equation (161) with $\beta_{VV} = \pi/2$ as

$$P_m = N/(V\theta_V)kT \cos^2 \theta_n \quad (165)$$

where $\theta_n = -\theta_\zeta$ and θ_ζ = phase angle of the internal space coordinate that is associated with particle number. As a first approximation $\theta_n \sim 0$ so that equation (165) becomes

$$P_m \sim N/(V\theta_V)kT \quad (166)$$

for the Boltzmann gas in coherent spacetime.

3. IDEAL FERMI GAS IN BROKEN SYMMETRY SPACETIME. This section calculates the chemical potential, energy density and pressure of an ideal Fermi gas that is located in a gravitational field or other external field which breaks the local symmetry of the spacetime coordinates. The cases of partially coherent and totally coherent spacetime are considered. These calculations may be of value for the study of the thermonuclear reactions in stars where spin 1/2 nuclei are the dominant reacting particle species as suggested in Section 5.

A. Complex Number Chemical Potential for Fermi Gas.

The chemical potential is determined by writing an expression for the total number of particles in terms of a sum over the complex number single particle momenta or single particle kinetic energies. The complex number generalization of the standard scalar result is easily written as³³⁻³⁵

$$\begin{aligned}\bar{n} &= \int_{\bar{p}} [1/\bar{z} \exp(\beta \bar{\epsilon}_{\bar{p}}) + 1]^{-1} \\ &= 4\pi V' / h^3 \int \bar{p}^2 \{1/\bar{z} \exp[\beta \bar{p}^2 / (2m)] + 1\}^{-1} d\bar{p} \\ &= fV' \int \bar{\epsilon}^{1/2} [1/\bar{z} \exp(\beta \bar{\epsilon}) + 1]^{-1} d\bar{\epsilon} \\ &= fV' \int \bar{\epsilon}^{1/2} [\exp(\bar{\xi}) + 1]^{-1} d\bar{\epsilon}\end{aligned}\tag{167}$$

where f is given by equation (77) and where

$$\bar{\xi} = \bar{\beta}(\bar{\epsilon} - \bar{\mu})\tag{168}$$

where \bar{z} and $\bar{\mu}$ are related by equation (70). The following series expansion is used to evaluate the integral in equation (167)

$$\begin{aligned}[\exp(\bar{\xi}) + 1]^{-1} &= \exp(-\bar{\xi}) [1 + \exp(-\bar{\xi})]^{-1} \\ &= \sum_{\nu=1}^{\infty} (-1)^{\nu-1} \exp(-\nu \bar{\xi}) \\ &= \sum_{\nu=1}^{\infty} (-1)^{\nu-1} \bar{z}^{\nu} \exp(-\nu \beta \bar{\epsilon})\end{aligned}\tag{169}$$

Then equation (167) can be written as

$$\bar{n} = fV' \sum_{\nu=1}^{\infty} (-1)^{\nu-1} \bar{z}^{\nu} \bar{H}_{\nu}\tag{170}$$

where

$$\bar{H}_{\nu} = H_{\nu} \exp(j\theta_{H\nu}) = H_{\nu R} + jH_{\nu I} = \int \bar{\epsilon}^{1/2} \exp(-\nu \beta \bar{\epsilon}) d\bar{\epsilon}\tag{171}$$

and where f is given by equation (77).

The integral in equation (171) can be rewritten for partially coherent spacetime as

$$\bar{H}_V = \int_0^{\infty} \epsilon^{1/2} \sec \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_{\epsilon}) \exp[j(3/2\theta_{\epsilon} - v\beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon})] d\epsilon \quad (172)$$

$$= \int_0^{\pi/3} \epsilon^{3/2} \csc \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_{\epsilon}) \exp[j(3/2\theta_{\epsilon} - v\beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon})] d\theta_{\epsilon} \quad (173)$$

or in component form as

$$H_{VR} = \int_0^{\infty} \epsilon^{1/2} \sec \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_{\epsilon}) \cos(3/2\theta_{\epsilon} - v\beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\epsilon \quad (174)$$

$$= \int_0^{\pi/3} \epsilon^{3/2} \csc \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_{\epsilon}) \cos(3/2\theta_{\epsilon} - v\beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\theta_{\epsilon} \quad (175)$$

$$H_{VI} = \int_0^{\infty} \epsilon^{1/2} \sec \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_{\epsilon}) \sin(3/2\theta_{\epsilon} - v\beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\epsilon \quad (176)$$

$$= \int_0^{\pi/3} \epsilon^{3/2} \csc \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_{\epsilon}) \sin(3/2\theta_{\epsilon} - v\beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon}) d\theta_{\epsilon} \quad (177)$$

These component terms give

$$H_V = (H_{VR}^2 + H_{VI}^2)^{1/2} \quad \tan \theta_{H_V} = H_{VI}/H_{VR} \quad (178)$$

and then equation (179) can be written as

$$\bar{n} = fV' \sum_{v=1}^{\infty} (-1)^{v-1} z^v H_V \exp[j(v\theta_z + \theta_{H_V})] \quad (179)$$

Taking the real and imaginary parts of equation (179) and using equations (64) and (65) gives

$$N \cos^2 \theta_n = fV' \sum_{v=1}^{\infty} (-1)^{v-1} z^v H_V \cos(v\theta_z + \theta_{H_V}) \quad (180)$$

$$N \cos \theta_n \sin \theta_n = fV' \sum_{v=1}^{\infty} (-1)^{v-1} z^v H_V \sin(v\theta_z + \theta_{H_V}) \quad (181)$$

where N = integer number of fermions. Equations (180) and (181) are two equations which can be used to determine z and θ_z from which μ and θ_{μ} can be obtained by using equations (73) and (74). Equation (180) gives the measured particle number as shown in equation (66).

The case of a Fermi gas in incoherent spacetime is obtained by taking all internal phase angles to be equal to zero. Then equation (170) gives

$$N = fV \sum_{v=1}^{\infty} (-1)^{v-1} z_{inc}^v H_v^{inc} \quad (182)$$

where from equation (183)

$$\begin{aligned} H_V^{inc} &= \int_0^\infty \epsilon^{1/2} \exp(-v\beta\epsilon) d\epsilon \\ &= (v\beta)^{-3/2} \Gamma(3/2) = 1/2\sqrt{\pi} (v\beta)^{-3/2} = 1/2\sqrt{\pi} v^{-3/2} (kT)^{3/2} \end{aligned} \quad (183)$$

which is the standard result for the ideal Fermi gas.³³⁻³⁵ Equation (182) can be used to determine z_{inc} . Following the same arguments given in equations (94) through (103) it is easy to show that for the case $\theta_\epsilon = \theta_\epsilon^1 = \text{constant}$ equations (174) and (176) become

$$H_{VR}^{inc} = 1/2\sqrt{\pi} (v\beta)^{-3/2} \quad (184)$$

$$H_{VI}^{inc} = 0 \quad (185)$$

which are independent of the angle θ_ϵ^1 , and which are identical to equation (183) which is valid for $\theta_\epsilon = 0$.

The particle number for fermions in coherent spacetime can be obtained by taking $V' = V\theta_V$ in equation (170) and by taking $\beta_{\epsilon\epsilon} = \pi/2$ in equations (173) and (175) with the result

$$\bar{n} = fV\theta_V \sum_{v=1}^\infty (-1)^{v-1} z_{coh}^v \bar{H}_v^{coh} \quad (186)$$

$$= fV\theta_V \sum_{v=1}^\infty (-1)^{v-1} z_{coh}^v H_v^{coh} \exp[j(v\theta_z^{coh} + \theta_{Hv}^{coh})] \quad (187)$$

where

$$\bar{H}_v^{coh} = \epsilon_c^{3/2} \bar{E}_v = \epsilon_c^{3/2} E_v \exp(j\theta_{Ev}) = \epsilon_c^{3/2} (E_{VR} + jE_{VI}) \quad (188)$$

$$H_v^{coh} = \epsilon_c^{3/2} E_v \quad \theta_{Hv}^{coh} = \theta_{Ev} \quad (189)$$

where $\epsilon_c = \text{constant}$ magnitude of the single particle kinetic energy, and where

$$\bar{E}_v = j \int_0^{\pi/3} \exp(j3/2\theta_\epsilon) \exp(-v\beta\bar{E}) d\theta_\epsilon \quad (190)$$

or

$$\bar{E}_v = j \int_0^{\pi/3} \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \exp[j(3/2\theta_\epsilon - v\beta\epsilon_c \sin \theta_\epsilon)] d\theta_\epsilon \quad (191)$$

which can be written in component form as

$$E_{VR} = \int_0^{\pi/3} \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \sin(v\beta\epsilon_c \sin \theta_\epsilon - 3/2\theta_\epsilon) d\theta_\epsilon \quad (192)$$

$$E_{VI} = \int_0^{\pi/3} \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \cos(v\beta\epsilon_c \sin \theta_\epsilon - 3/2\theta_\epsilon) d\theta_\epsilon \quad (193)$$

$$E_v = (E_{vR}^2 + E_{vI}^2)^{1/2} \quad \tan \theta_{Ev} = E_{vI}/E_{vR} \quad (194)$$

The values of $\bar{z}_{coh} = (z_{coh}, \theta_z^{coh})$ can be obtained from equations (186) through (189) by writing

$$\bar{n} = fV\theta_v \epsilon_c^{3/2} \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{coh}^v \bar{E}_v \quad (195)$$

or equivalently as

$$\bar{n} = fV\theta_v \epsilon_c^{3/2} \sum_{v=1}^{\infty} (-1)^{v-1} z_{coh}^v E_v \exp[j(v\theta_z^{coh} + \theta_{Ev})] \quad (196)$$

The real and imaginary parts of equation (196) give

$$N \cos^2 \theta_n = fV\theta_v \epsilon_c^{3/2} \sum_{v=1}^{\infty} (-1)^{v-1} z_{coh}^v E_v \cos(v\theta_z^{coh} + \theta_{Ev}) \quad (197)$$

$$N \cos \theta_n \sin \theta_n = fV\theta_v \epsilon_c^{3/2} \sum_{v=1}^{\infty} (-1)^{v-1} z_{coh}^v E_v \sin(v\theta_z^{coh} + \theta_{Ev}) \quad (198)$$

Equations (197) and (198) are the two equations that determine z_{coh} and θ_z^{coh} for coherent spacetime. Equations (73) and (74) can be used to determine μ_{coh} and θ_μ^{coh} . Equation (197) is the expression for the measured fermion number for coherent spacetime in accordance with equation (66).

B. Internal Energy of Fermi Gas.

For a Fermi gas located in a strong gravitational or magnetic field the spacetime symmetry is broken and the internal energy is written as a generalization of the standard scalar form as follows³³⁻³⁵

$$\bar{U} = \bar{U} \exp(j\theta_U) = fV' \int \bar{\epsilon}^{3/2} [\exp(\bar{\xi}) + 1]^{-1} d\bar{\epsilon} \quad (199)$$

where f and $\bar{\xi}$ are given by equations (77) and (168) respectively. Using the series expansion in equation (169) gives

$$\bar{U} = fV' \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}^v \bar{J}_v \quad (200)$$

$$= fV' \sum_{v=1}^{\infty} (-1)^{v-1} z^v J_v \exp[j(v\theta_z + \theta_{Jv})] \quad (201)$$

where

$$\bar{J}_v = J_v \exp(j\theta_{Jv}) = J_{vR} + jJ_{vI} = \int \bar{\epsilon}^{3/2} \exp(-v\beta\bar{\epsilon}) d\bar{\epsilon} \quad (202)$$

$$J_v = (J_{vR}^2 + J_{vI}^2)^{1/2} \quad \tan \theta_{Jv} = J_{vI}/J_{vR} \quad (203)$$

The problem then is to determine J_{vR} and J_{vI} .

For partially coherent spacetime the integral in equation (202) can be written as

$$\bar{J}_V = \int_0^\infty \epsilon^{3/2} \sec \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_\epsilon) \exp[j(5/2\theta_\epsilon - v\beta\epsilon \sin \theta_\epsilon + \beta_{\epsilon\epsilon})] d\epsilon \quad (204)$$

$$= \int_0^{\pi/3} \epsilon^{5/2} \csc \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_\epsilon) \exp[j(5/2\theta_\epsilon - v\beta\epsilon \sin \theta_\epsilon + \beta_{\epsilon\epsilon})] d\theta_\epsilon \quad (205)$$

from which the real and imaginary parts are obtained as

$$J_{VR} = \int_0^\infty \epsilon^{3/2} \sec \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_\epsilon) \cos(5/2\theta_\epsilon - v\beta\epsilon \sin \theta_\epsilon + \beta_{\epsilon\epsilon}) d\epsilon \quad (206)$$

$$= \int_0^{\pi/3} \epsilon^{5/2} \csc \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_\epsilon) \cos(5/2\theta_\epsilon - v\beta\epsilon \sin \theta_\epsilon + \beta_{\epsilon\epsilon}) d\theta_\epsilon \quad (207)$$

$$J_{VI} = \int_0^\infty \epsilon^{3/2} \sec \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_\epsilon) \sin(5/2\theta_\epsilon - v\beta\epsilon \sin \theta_\epsilon + \beta_{\epsilon\epsilon}) d\epsilon \quad (208)$$

$$= \int_0^{\pi/3} \epsilon^{5/2} \csc \beta_{\epsilon\epsilon} \exp(-v\beta\epsilon \cos \theta_\epsilon) \sin(5/2\theta_\epsilon - v\beta\epsilon \sin \theta_\epsilon + \beta_{\epsilon\epsilon}) d\theta_\epsilon \quad (209)$$

These component values can be used to determine J_V and θ_{J_V} by equation (203), and then \bar{U} is calculated from equation (201) in terms of z and θ_z which are determined from equations (180) and (181). The real and imaginary components of the internal energy are obtained from equation (201) as

$$U_R = fV' \sum_{v=1}^{\infty} (-1)^{v-1} z^v J_V \cos(v\theta_z + \theta_{J_V}) \quad (210)$$

$$U_I = fV' \sum_{v=1}^{\infty} (-1)^{v-1} z^v J_V \sin(v\theta_z + \theta_{J_V}) \quad (211)$$

The measured value of the internal energy is given by $U_m = U_R$. The average complex number energy per fermion is obtained from equations (170) and (200) to be

$$\bar{\epsilon}_{av} = \bar{U}/\bar{n} = \left[\sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}^v \bar{J}_V \right] / \left[\sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}^v \bar{H}_V \right] \quad (212)$$

whose real part gives the measured average energy per fermion as $\epsilon_{avm} = \epsilon_{avR}$.

The internal energy for the case of incoherent spacetime is obtained by setting all internal phase angles equal to zero and equation (201) becomes

$$U^{inc} = fV' \sum_{v=1}^{\infty} (-1)^{v-1} z_{inc}^v J_V^{inc} \quad (213)$$

where from equation (202)

$$J_v^{inc} = \int_0^\infty \epsilon^{3/2} \exp(-v\beta\epsilon) d\epsilon \quad (214)$$

$$= (v\beta)^{-5/2} \Gamma(5/2) = 3/4\sqrt{\pi} v^{-5/2} (kT)^{5/2}$$

Equations (213) and (214) represent the standard case for the internal energy of an ideal Fermi gas.³³⁻³⁵ A comparison of equations (183) and (214) show that

$$H_v^{inc} = 2/3\beta v J_v^{inc} \quad (215)$$

For the incoherent spacetime case with $\theta_\epsilon = \theta_\epsilon^1 = \text{constant}$ it follows by the same arguments given in equations (130) through (141) that equations (206) and (208) become

$$J_{vR}^{inc} = 3/4\sqrt{\pi} (v\beta)^{-5/2} \quad (216)$$

$$J_{vI}^{inc} = 0 \quad (217)$$

which is identical to the case in equation (214) that was obtained by taking $\theta_\epsilon = 0$.

The case of coherent spacetime can be obtained by taking $\beta_{\epsilon\epsilon} = \pi/2$, $V' = V\theta_V$ and $\epsilon = \epsilon_c = \text{constant}$ so that equation (201) becomes

$$\bar{U}^{coh} = fV\theta_V \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{coh}^v J_v^{coh} \quad (218)$$

$$= fV\theta_V \sum_{v=1}^{\infty} (-1)^{v-1} z_{coh}^v J_v^{coh} \exp[j(v\theta_z^{coh} + \theta_{Jv}^{coh})] \quad (219)$$

where f is given by equation (77), z_{coh} and θ_z^{coh} are given by equations (197) and (198), and where for coherent spacetime equations (205), (207) and (209) become

$$\bar{J}_v^{coh} = \epsilon_c^{5/2} \bar{S}_v = \epsilon_c^{5/2} S_v \exp(j\theta_{Sv}) = \epsilon_c^{5/2} (S_{vR} + jS_{vI}) \quad (220)$$

or

$$J_v^{coh} = \epsilon_c^{5/2} S_v \quad \theta_{Jv}^{coh} = \theta_{Sv} \quad (221)$$

where

$$\bar{S}_v = j \int_0^{\pi/3} \exp(j5/2\theta_\epsilon) \exp(-v\beta\epsilon) d\theta_\epsilon \quad (222)$$

$$= j \int_0^{\pi/3} \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \exp[j(5/2\theta_\epsilon - v\beta\epsilon_c \sin \theta_\epsilon)] d\theta_\epsilon \quad (223)$$

$$S_{vR} = \int_0^{\pi/3} \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \sin(v\beta\epsilon_c \sin \theta_\epsilon - 5/2\theta_\epsilon) d\theta_\epsilon \quad (224)$$

$$S_{VI} = \int_0^{\pi/3} \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \cos(v\beta\epsilon_c \sin \theta_\epsilon - 5/2\theta_\epsilon) d\theta_\epsilon \quad (225)$$

and

$$S_v = (S_{vR}^2 + S_{vI}^2)^{1/2} \quad \tan \theta_{Sv} = S_{vI}/S_{vR} \quad (226)$$

Therefore the internal energy for fermions in coherent spacetime can be written as

$$\bar{U}^{\text{coh}} = fv\theta_{v\epsilon_c}^{5/2} \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v \bar{S}_v \quad (227)$$

so that

$$U_R^{\text{coh}} = fv\theta_{v\epsilon_c}^{5/2} \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v S_v \cos(v\theta_z^{\text{coh}} + \theta_{Sv}) \quad (228)$$

$$U_I^{\text{coh}} = fv\theta_{v\epsilon_c}^{5/2} \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v S_v \sin(v\theta_z^{\text{coh}} + \theta_{Sv}) \quad (229)$$

The measured value of the internal energy in coherent spacetime is given by equation (228), and can be a negative quantity. The average value of the internal energy per fermion in the coherent spacetime state is obtained from equations (195) and (227) to be

$$\begin{aligned} \bar{\epsilon}^{\text{coh}} &= \bar{U}^{\text{coh}}/\bar{n} \\ &= \epsilon_c \left[\sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v \bar{S}_v \right] / \left[\sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v \bar{E}_v \right] \end{aligned} \quad (230)$$

where \bar{S}_v and \bar{E}_v are given by equations (222) and (190) respectively.

C. Pressure of a Fermi Gas in a Gravitational Field.

The expression for the pressure of a Fermi gas located in broken symmetry spacetime is given by the following complex number generalization of the standard statistical mechanics expression³³⁻³⁵

$$\bar{P}/(kT) = 1/V' \sum_p \ln[1 + \bar{z} \exp(-\beta\bar{\epsilon}_p)] \quad (231)$$

$$= 4\pi/h^3 \int \bar{p}^2 \ln[1 + \bar{z} \exp(-\beta\bar{\epsilon}_p)] d\bar{p} \quad (232)$$

$$= 2\pi(2m)^{3/2}/h^3 \int \bar{\epsilon}^{1/2} \ln[1 + \bar{z} \exp(-\beta\bar{\epsilon})] d\bar{\epsilon} \quad (233)$$

It is easy to verify the following series expansion

$$\ln[1 + \bar{z} \exp(-\beta\bar{\epsilon})] = \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}^v/v \exp(-v\beta\bar{\epsilon}) \quad (234)$$

Then the pressure of a Fermi gas in partially coherent spacetime is given by

$$\bar{P}/(kT) = f \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}^v / v \bar{H}_v \quad (235)$$

where f and \bar{H}_v are given by equations (77) and (171) respectively. The result for the pressure in equation (235) can be obtained directly by combining equation (170) with the completely general equation (154). From the general relation for the ideal gases

$$\bar{P} = P \exp(j\theta_p) = 2/3\bar{U}/v' \quad (236)$$

it follows from equations (200) and (235) that

$$\bar{H}_v = 2/3\beta v \bar{J}_v \quad (237)$$

Equation (235) can also be written as

$$\bar{P}/(kT) = f \sum_{v=1}^{\infty} (-1)^{v-1} z^v / v H_v \exp[j(v\theta_z + \theta_{Hv})] \quad (238)$$

so that

$$P_R/(kT) = f \sum_{v=1}^{\infty} (-1)^{v-1} z^v / v H_v \cos(v\theta_z + \theta_{Hv}) \quad (239)$$

$$P_I/(kT) = f \sum_{v=1}^{\infty} (-1)^{v-1} z^v / v H_v \sin(v\theta_z + \theta_{Hv}) \quad (240)$$

and

$$\tan \theta_p = P_I/P_R \quad (241)$$

The measured pressure is given by equation (239).

For the incoherent spacetime case the pressure of a Fermi gas is obtained from equation (235) as

$$\begin{aligned} P^{inc} &= f kT \sum_{v=1}^{\infty} (-1)^{v-1} z_{inc}^v / v H_v^{inc} \\ &= 2/3f \sum_{v=1}^{\infty} (-1)^{v-1} z_{inc}^v J_v^{inc} \end{aligned} \quad (242)$$

where H_v^{inc} is given by equations (183) and (215), and

$$J_v^{inc} = H_v^{inc} / (2/3\beta v) = 3/4\sqrt{\pi} v^{-5/2} (kT)^{5/2} \quad (243)$$

which is the standard textbook result.³³⁻³⁵

For coherent spacetime the pressure of a Fermi gas can be obtained from equations (188) and (235) to be

$$\bar{p}^{\text{coh}}/(kT) = f \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v / v \bar{H}_v^{\text{coh}} \quad (244)$$

$$= f \epsilon_c^{3/2} \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v / v \bar{E}_v \quad (245)$$

$$= f \epsilon_c^{3/2} \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v / v E_v \exp[j(\theta_z^{\text{coh}} + \theta_{E_v})] \quad (246)$$

For a coherent spacetime the pressure is given by

$$\bar{p}^{\text{coh}} = 2/3 \bar{U}^{\text{coh}} / (v \theta_v) \quad (247)$$

and a comparison of equations (218), (220), (227), (245) and (247) gives

$$\bar{H}_v^{\text{coh}} = 2/3 \beta v \bar{J}_v^{\text{coh}} \quad (248)$$

or

$$\bar{E}_v = 2/3 \beta v \epsilon_c \bar{S}_v \quad (249)$$

or equivalently

$$E_v = 2/3 \beta v \epsilon_c S_v \quad \theta_{E_v} = \theta_{S_v} \quad (250)$$

Equivalently the coherent spacetime Fermi gas pressure is written as

$$\bar{p}^{\text{coh}} = 2/3 f \epsilon_c^{5/2} \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v \bar{S}_v \quad (251)$$

$$= 2/3 f \epsilon_c^{5/2} \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v S_v \exp[j(v \theta_z^{\text{coh}} + \theta_{S_v})] \quad (252)$$

Finally, the real and imaginary parts of the Fermi gas pressure in coherent spacetime is given by

$$p_R^{\text{coh}} = f k T \epsilon_c^{3/2} \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v / v E_v \cos(v \theta_z^{\text{coh}} + \theta_{E_v}) \quad (253)$$

$$= 2/3 f \epsilon_c^{5/2} \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v S_v \cos(v \theta_z^{\text{coh}} + \theta_{E_v}) \quad (254)$$

$$p_I^{\text{coh}} = f k T \epsilon_c^{3/2} \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v / v E_v \sin(v \theta_z^{\text{coh}} + \theta_{E_v}) \quad (255)$$

$$= 2/3 f \epsilon_c^{5/2} \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v S_v \sin(v \theta_z^{\text{coh}} + \theta_{E_v}) \quad (256)$$

where S_v and θ_{S_v} are given by equation (226), and E_v and θ_{E_v} are given by equation (194).

D. Zero Temperature Fermi Gas in the Presence of Gravity.

Consider now the $T = 0$ state of an ideal Fermi gas in spacetime with broken internal symmetries. The complex number generalization of the standard scalar expression for the particle number is given by³³⁻³⁵

$$\bar{n} = 4\pi V'/h^3 \int \bar{p}^2 d\bar{p} \quad (257)$$

$$= 4\pi V'/h^3 \int_0^\infty p^2 \sec \beta_{pp} \exp[j(3\theta_p + \beta_{pp})] dp \quad (258)$$

$$= 4\pi V'/h^3 \int_0^{\pi/6} p^3 \csc \beta_{pp} \exp[j(3\theta_p + \beta_{pp})] d\theta_p \quad (259)$$

where β_{pp} is given by equation (49). Equivalently these equations can be written in terms of single particle energy as

$$\bar{n} = fV' \int \bar{\epsilon}^{1/2} d\bar{\epsilon} \quad (260)$$

$$= fV' \int_0^\infty \epsilon^{1/2} \sec \beta_{\epsilon\epsilon} \exp[j(3/2\theta_\epsilon + \beta_{\epsilon\epsilon})] d\epsilon \quad (261)$$

$$= fV' \int_0^{\pi/3} \epsilon^{3/2} \csc \beta_{\epsilon\epsilon} \exp[j(3/2\theta_\epsilon + \beta_{\epsilon\epsilon})] d\theta_\epsilon \quad (262)$$

where V' , $\beta_{\epsilon\epsilon}$ and f are given by equations (6), (54) and (77). For incoherent spacetime these integrals reduce to the standard expressions for the $T = 0$ Fermi gas. For coherent spacetime with $\beta_{pp} = \pi/2$, $\beta_{\epsilon\epsilon} = \pi/2$, $V' = V\theta_V$, $p = p_c = \text{constant}$, and $\epsilon = \epsilon_c = \text{constant}$ it follows from equations (257) and (260) or (259) and (262) that for internal space rotations of the momentum and single particle energy

$$\bar{n} = 4/3\pi V\theta_V/h^3 \bar{p}^3 \Big|_{p_c}^{\bar{p}_c} = 4/3\pi V\theta_V/h^3 (\bar{p}_c^3 - p_c^3) \quad (263)$$

$$= 2/3fV\theta_V \bar{\epsilon}^{3/2} \Big|_{\epsilon_c}^{\bar{\epsilon}_c} = 2/3fV\theta_V (\bar{\epsilon}_c^{3/2} - \epsilon_c^{3/2}) \quad (264)$$

where

$$\bar{p}_c = p_c \exp(j\pi/6) \quad (265)$$

$$\bar{\epsilon}_c = \epsilon_c \exp(j\pi/3) \quad (266)$$

Combining equations (263) through (266) gives for the $T = 0$ Fermi gas in coherent spacetime

$$\bar{n} = 4/3\pi V\theta_V p_c^3 (j - 1)/h^3 \quad (267)$$

$$= 4/3\pi V\theta_V (2m\epsilon_c)^{3/2} (j - 1)/h^3 \quad (268)$$

$$= 2/3fV\theta_V \epsilon_c^{3/2} (j - 1) \quad (269)$$

Note that for this case $n_R < 0$. For particles with spin the above results must be multiplied by $(2s + 1)$ where $s = \text{spin}$.

The internal energy of a $T = 0$ Fermi gas in the presence of gravity is written as the following complex number generalization of the standard scalar result³³⁻³⁵

$$\bar{U} = 4\pi V' / (2mh^3) \int \bar{p}^4 d\bar{p} \quad (270)$$

$$= 4\pi V' / (2mh^3) \int_0^\infty p^4 \sec \beta_{pp} \exp[j(5\theta_p + \beta_{pp})] dp \quad (271)$$

$$= 4\pi V' / (2mh^3) \int_0^{\pi/6} p^5 \csc \beta_{pp} \exp[j(5\theta_p + \beta_{pp})] d\theta_p \quad (272)$$

These equations can be written equivalently in terms of the single particle energy as

$$\bar{U} = fV' \int \bar{\epsilon}^{3/2} d\bar{\epsilon} \quad (273)$$

$$= fV' \int_0^\infty \epsilon^{3/2} \sec \beta_{\epsilon\epsilon} \exp[j(5/2\theta_\epsilon + \beta_{\epsilon\epsilon})] d\epsilon \quad (274)$$

$$= fV' \int_0^{\pi/3} \epsilon^{5/2} \csc \beta_{\epsilon\epsilon} \exp[j(5/2\theta_\epsilon + \beta_{\epsilon\epsilon})] d\theta_\epsilon \quad (275)$$

For incoherent spacetime all internal phase angles are set equal to zero and equations (270), (271), (273) and (274) reduce to the standard scalar expressions. For coherent spacetime with $\beta_{pp} = \pi/2$, $\beta_{\epsilon\epsilon} = \pi/2$, $V' = V\theta_V$, $p = p_c$ = constant, and $\epsilon = \epsilon_c$ = constant it follows from equations (270) and (273) or equations (272) and (275) that for the case of internal spacetime rotations of the single particle momentum and energy

$$\bar{U} = 2/5\pi V\theta_V / (mh^3) \bar{p}^5|_{p_c} = 2/5\pi V\theta_V / (mh^3) (\bar{p}_c^5 - p_c^5) \quad (276)$$

$$= 2/5fV\theta_V \bar{\epsilon}^{5/2}|_{\epsilon_c} = 2/5fV\theta_V (\bar{\epsilon}_c^{5/2} - \epsilon_c^{5/2}) \quad (277)$$

Combining equations (265), (266), (276) and (277) gives

$$\bar{U} = 1/5\pi V\theta_V p_c^5 / (mh^3) (j - \sqrt{3} - 2) \quad (278)$$

$$= 1/5fV\theta_V \epsilon_c^{5/2} (j - \sqrt{3} - 2) \quad (279)$$

For this case $U_R < 0$. For particles with spin these results must be multiplied by $(2s + 1)$ where $s = \text{spin}$. The average energy per particle is obtained from equations (267), (269), (278) and (279) to be

$$\bar{\epsilon}_{av} = \bar{U}/\bar{n} \quad (280)$$

$$= 3/10\epsilon_c (j - \sqrt{3} - 2)/(j - 1) \quad (281)$$

$$= 3/10(1 + \sqrt{3})\epsilon_c \exp(j\pi/6) \quad (282)$$

The real part of equation (282) is given by

$$\epsilon_{avR} = 3\sqrt{3}(1 + \sqrt{3})/20 \epsilon_c \quad (283)$$

$$\sim 0.71\epsilon_c$$

The result in equation (283) can be compared with the corresponding result for incoherent spacetime which is $\epsilon_{av} = 3/5\epsilon_F = 0.6\epsilon_F$ where ϵ_F = Fermi energy.

4. BOSE GAS IN BROKEN SYMMETRY SPACETIME. This section calculates the chemical potential, internal energy and pressure of a Bose gas which is located in an external field such as gravity or an electromagnetic field which breaks the symmetry of spacetime and induces complex number values for the spacetime coordinates, particle momenta and single particle energies. This requires that the total particle number, chemical potential, internal energy and pressure must be represented as complex numbers.

A. Chemical Potential of Bose Gas.

For a Bose gas in a gravitational field the complex number generalization of the standard expression for the total particle number in a Bose gas is given by

$$\bar{n} = \sum_p [1/\bar{z} \exp(\beta\bar{\epsilon}_p) - 1]^{-1} \quad (284)$$

$$= \sum_{p \neq 0} \sum_{v=1}^{\infty} \bar{z}^v \exp(-v\bar{\epsilon}_p) + \bar{z}/(1 - \bar{z}) \quad (285)$$

$$= fV' \sum_{v=1}^{\infty} \bar{z}^v \bar{H}_v + \bar{z}/(1 - \bar{z}) \quad (286)$$

where V' , f and \bar{H}_v are given by equations (6), (77) and (171) respectively. Equation (286) can be rewritten as

$$\bar{n} = fV' \sum_{v=1}^{\infty} z^v H_v \exp[j(v\theta_z + \theta_{Hv})] + \bar{z}/(1 - \bar{z}) \quad (287)$$

where the fugacity \bar{z} is defined in equation (70). The real and imaginary parts of equation (287) are obtained using equations (64) and (65) as

$$\begin{aligned} N \cos^2 \theta_n &= fV' \sum_{v=1}^{\infty} z^v H_v \cos(v\theta_z + \theta_{Hv}) \\ &+ z(\cos \theta_z - z)/(1 + z^2 - 2z \cos \theta_z) \end{aligned} \quad (288)$$

$$N \cos \theta_n \sin \theta_n = fV' \sum_{v=1}^{\infty} z^v H_v \sin(v\theta_z + \theta_{Hv}) \quad (289)$$

$$+ (z \sin \theta_z)/(1 + z^2 - 2z \cos \theta_z)$$

where N = integer number of bosons. Equations (288) and (289) can be used to determine z and θ_z from which the chemical potential components μ and θ_μ are obtained from equations (73) and (74). From equation (66) it follows that equation (288) gives the measured particle number.

For incoherent spacetime all internal phase angles are set equal to zero and equation (288) becomes

$$N = fV \sum_{v=1}^{\infty} z_{inc}^v H_v^{inc} + z_{inc}/(1 - z_{inc}) \quad (290)$$

where H_v^{inc} is given by equation (183), and z_{inc} is obtained as a solution of equation (290). Equation (290) is the standard textbook expression for the particle number for the case of an ideal Bose gas.³³⁻³⁵

For the case of a Bose gas in coherent spacetime with $V' = V\theta_V$, $\beta_{\epsilon\epsilon} = \pi/2$ and $\epsilon = \epsilon_c = \text{constant}$, it is easy to show that

$$\bar{n} = fV\theta_V \sum_{v=1}^{\infty} \bar{z}_{coh}^v \bar{H}_v^{coh} + \bar{z}_{coh}/(1 - \bar{z}_{coh}) \quad (291)$$

$$= fV\theta_V \epsilon_c^{3/2} \sum_{v=1}^{\infty} \bar{z}_{coh}^v \bar{E}_v + \bar{z}_{coh}/(1 - \bar{z}_{coh}) \quad (292)$$

where \bar{E}_v is given by equation (191). Then it follows that

$$N \cos^2 \theta_n = fV\theta_V \epsilon_c^{3/2} \sum_{v=1}^{\infty} z_{coh}^v E_v \cos(v\theta_z^{coh} + \theta_{Ev}) \quad (293)$$

$$+ z_{coh} (\cos \theta_z^{coh} - z_{coh})/(1 + z_{coh}^2 - 2z_{coh} \cos \theta_z^{coh})$$

$$N \cos \theta_n \sin \theta_n = fV\theta_V \epsilon_c^{3/2} \sum_{v=1}^{\infty} z_{coh}^v E_v \sin(v\theta_z^{coh} + \theta_{Ev}) \quad (294)$$

$$+ (z_{coh} \sin \theta_z^{coh})/(1 + z_{coh}^2 - 2z_{coh} \cos \theta_z^{coh})$$

where E_v and θ_{Ev} are given by equation (194). Equations (293) and (294) can be used to determine z_{coh} and θ_z^{coh} and subsequently μ_{coh} and θ_μ^{coh} .

B. Internal Energy of Bose Gas in a Gravity Field.

The internal energy of a Bose gas in broken symmetry spacetime is written as the following complex number generalization of the scalar result³³⁻³⁵

$$\bar{U} = fV' \int \bar{\epsilon}^{3/2} [\exp(\bar{\epsilon}) - 1]^{-1} d\bar{\epsilon} - 3/2kT \ln(1 - \bar{z}) \quad (295)$$

where $\bar{\xi}$ is given by equation (168). Expanding the integrand by an infinite series gives

$$\bar{U} = fV' \sum_{v=1}^{\infty} \bar{z}^v \bar{J}_v - 3/2kT \ln(1 - \bar{z}) \quad (296)$$

where \bar{J}_v is given by equation (202). The real and imaginary parts of equation (296) are obtained to be

$$U_R = fV' \sum_{v=1}^{\infty} z^v J_v \cos(v\theta_z + \theta_{Jv}) - 3/2kT \ln Y \quad (297)$$

$$U_I = fV' \sum_{v=1}^{\infty} z^v J_v \sin(v\theta_z + \theta_{Jv}) + 3/2kT\theta_Y \quad (298)$$

where J_v and θ_{Jv} are given by equations (203), and where

$$1 - \bar{z} = Y \exp(-j\theta_Y) \quad (299)$$

$$Y = (1 - 2z \cos \theta_z + z^2)^{1/2} \quad (300)$$

$$\tan \theta_Y = (z \sin \theta_z) / (1 - z \cos \theta_z) \quad (301)$$

The measured internal energy is given by $U_m = U_R$. The average energy per particle for the Bose gas in a gravity field is given by

$$\bar{\epsilon}_{av} = \bar{U}/\bar{n} \quad (302)$$

where \bar{U} is given by equation (296) and \bar{n} is given by equation (286).

For incoherent spacetime all internal phase angles have zero values and equations (297) and (299) give the internal energy of an ideal Bose gas as

$$U_{inc} = fV \sum_{v=1}^{\infty} z_{inc}^v J_v^{inc} - 3/2kT \ln(1 - z_{inc}) \quad (303)$$

where J_v^{inc} is given by equation (214), and z_{inc} is determined as a solution to equation (290). Equation (303) represents the standard expression for the internal energy of an ideal Bose gas.³³⁻³⁵

For coherent spacetime the internal energy is written as

$$\bar{U}^{coh} = fV\theta_V \sum_{v=1}^{\infty} \bar{z}_{coh}^v \bar{J}_v^{coh} - 3/2kT \ln(1 - \bar{z}_{coh}) \quad (304)$$

$$= fV\theta_V \epsilon_c^{5/2} \sum_{v=1}^{\infty} \bar{z}_{coh}^v \bar{S}_v - 3/2kT \ln(1 - \bar{z}_{coh}) \quad (305)$$

where \bar{J}_v^{coh} and \bar{S}_v are defined by equations (220) through (222). The real and imaginary parts of the internal energy of a Bose gas located in a strong gravitational field is given by

$$U_R^{\text{coh}} = f v \theta_V \epsilon_c^{5/2} \sum_{v=1}^{\infty} z_{\text{coh}}^v S_v \cos(v \theta_z^{\text{coh}} + \theta_{Sv}) - 3/2 kT \ln Y_{\text{coh}} \quad (306)$$

$$U_I^{\text{coh}} = f v \theta_V \epsilon_c^{5/2} \sum_{v=1}^{\infty} z_{\text{coh}}^v S_v \sin(v \theta_z^{\text{coh}} + \theta_{Sv}) + 3/2 kT \theta_Y^{\text{coh}} \quad (307)$$

where from equations (299) through (301)

$$1 - \bar{z}_{\text{coh}} = Y_{\text{coh}} \exp(-j \theta_Y^{\text{coh}}) \quad (308)$$

$$Y_{\text{coh}} = (1 - 2 z_{\text{coh}} \cos \theta_z^{\text{coh}} + z_{\text{coh}}^2)^{1/2} \quad (309)$$

$$\tan \theta_Y^{\text{coh}} = (z_{\text{coh}} \sin \theta_z^{\text{coh}}) / (1 - z_{\text{coh}} \cos \theta_z^{\text{coh}}) \quad (310)$$

The measured internal energy is given by $U_m^{\text{coh}} = U_R^{\text{coh}}$.

C. Pressure of a Bose Gas in a Gravity Field.

The pressure of an ideal Bose gas in a gravitational field can be immediately obtained from equations (236) and (296) to be

$$\bar{P} = 2/3 f \sum_{v=1}^{\infty} \bar{z}^v \bar{J}_v - 1/V' kT \ln(1 - \bar{z}) \quad (311)$$

This expression can also be obtained from the following complex number generalization of the standard scalar result for the pressure of an ideal Bose gas³³⁻³⁵

$$\bar{P}/(kT) = -f \int \bar{\epsilon}^{1/2} \ln[1 - \bar{z} \exp(-\beta \bar{\epsilon})] d\bar{\epsilon} - 1/V' \ln(1 - \bar{z}) \quad (312)$$

Then using the expression

$$\ln(1 - \bar{x}) = \sum_{v=1}^{\infty} (-1)^{v-1} \bar{x}^v / v \quad (313)$$

gives

$$\bar{P}/(kT) = f \sum_{v=1}^{\infty} \bar{z}^v / v \bar{H}^v - 1/V' \ln(1 - \bar{z}) \quad (314)$$

where \bar{H}^v is given by equation (171). A comparison of equations (311) and (314) shows that equation (237) is true also for the ideal Bose gas. The real and imaginary components of the pressure are obtained from equation (311) to be

$$P_R = 2/3 f \sum_{v=1}^{\infty} z^v J_v \cos(v \theta_z + \theta_{Jv}) - 1/V' kT \ln Y \quad (315)$$

$$P_I = 2/3 f \sum_{v=1}^{\infty} z^v J_v \sin(v \theta_z + \theta_{Jv}) + 1/V' kT \theta_Y \quad (316)$$

The measured pressure is given by $P_m = P_R$.

For an ideal Bose gas in incoherent spacetime all of the internal phase angles are set equal to zero and equation (315) becomes

$$P_{\text{inc}} = 2/3f \sum_{v=1}^{\infty} z_{\text{inc}}^v J_v^{\text{inc}} - V^{-1} kT \ln(1 - z_{\text{inc}}) \quad (317)$$

where J_v^{inc} is given by equation (243). Equation (317) is the standard expression for the pressure and in fact from equation (303) it follows that³³⁻³⁵

$$P_{\text{inc}} = 2/3U_{\text{inc}}/V \quad (318)$$

which agrees with equation (236).

The pressure of an ideal Bose gas in coherent spacetime is obtained from equation (236) to be

$$\bar{P}^{\text{coh}} = 2/3\bar{U}^{\text{coh}}/(V\theta_V) \quad (319)$$

$$= 2/3f\epsilon_c^{5/2} \sum_{v=1}^{\infty} \bar{z}_{\text{coh}}^v \bar{S}_v - (V\theta_V)^{-1} kT \ln(1 - \bar{z}_{\text{coh}}) \quad (320)$$

or from equations (306) and (307)

$$P_R^{\text{coh}} = 2/3f\epsilon_c^{5/2} \sum_{v=1}^{\infty} z_{\text{coh}}^v S_v \cos(v\theta_z^{\text{coh}} + \theta_{Sv}) - (V\theta_V)^{-1} kT \ln Y_{\text{coh}} \quad (321)$$

$$P_I^{\text{coh}} = 2/3f\epsilon_c^{5/2} \sum_{v=1}^{\infty} z_{\text{coh}}^v S_v \sin(v\theta_z^{\text{coh}} + \theta_{Sv}) + (V\theta_V)^{-1} kT\theta_Y^{\text{coh}} \quad (322)$$

where \bar{S}_v , S_v and θ_{Sv} are given in equations (222) and (226), and where Y_{coh} and θ_Y^{coh} are given by equations (309) and (310). The calculations in this section can be applied to the Bose gas of Cooper electron pairs in a high- T_c superconductor. The totally coherent spacetime state describes the superconducting state of a high- T_c material, whereas the partially coherent spacetime state can be used to describe the normal state of a high- T_c superconductor.

5. THERMONUCLEAR REACTIONS IN THE PRESENCE OF GRAVITY. This section calculates the rate of thermonuclear reactions of particles that exist in spacetime that has broken symmetries due to the presence of an external field such as gravity. The calculations will be of value for predicting the thermonuclear reaction rates in stars where the gravitational field is sufficiently strong as to make spacetime partially coherent or even totally coherent. The conventional calculation of the nuclear reaction rates in stars assumes that spacetime is incoherent. These conventional fusion rate calculations are treated extensively in the literature so that only a brief summary of the results are given here. The thermonuclear reaction rate R^{inc} is written as¹⁻⁶

$$R^{\text{inc}} = n_1 n_2 \langle \sigma v \rangle = n_1 n_2 / m \langle \sigma p \rangle = n_1 n_2 (2/m)^{1/2} \langle \sigma \epsilon^{1/2} \rangle \quad (323)$$

where σ = fusion reaction cross section, v = relative collision speed, m = re-

duced mass of the two reacting species of particles, $\epsilon = 1/2mv^2$ = kinetic energy of the two reacting particles in the center of mass system, and where n_1 and n_2 = nuclear species particle number densities given by

$$n_1 = N_1/V \quad n_2 = N_2/V \quad (324)$$

By way of introduction to the problem, the thermonuclear reaction rate is evaluated for an ideal Boltzmann gas in incoherent spacetime with a zero external gravitational field. For a Maxwell-Boltzmann distribution in incoherent spacetime the average value that appears in equation (323) is evaluated using equations (92) and (93) as follows¹⁻⁶

$$\begin{aligned} R^{inc} &= n_1 n_2 (2/m)^{1/2} F^{inc}/A^{inc} \\ &= n_1 n_2 [8/(\pi m)]^{1/2} (kT)^{-3/2} F^{inc} \\ &= 4\pi n_1 n_2 [m/(2\pi kT)]^{3/2} \int_0^\infty \exp(-1/2\beta m v^2) v^3 dv \end{aligned} \quad (325)$$

where A^{inc} is given by equation (93), and F^{inc} is given by

$$F^{inc} = \int_0^\infty \sigma \epsilon \exp(-\beta \epsilon) d\epsilon \quad (326)$$

where $\epsilon = 1/2mv^2$.

The generalization of the standard result in equation (323) to the case of broken symmetry spacetime that is induced by an external gravity field is written as the following complex number equation

$$\bar{R} = \bar{n}_1 \bar{n}_2 \langle \bar{\sigma} \bar{v} \rangle = \bar{n}_1 \bar{n}_2 / m \langle \bar{\sigma} \bar{p} \rangle = \bar{n}_1 \bar{n}_2 (2/m)^{1/2} \langle \bar{\sigma} \bar{\epsilon} \rangle^{1/2} \quad (327)$$

where \bar{R} = complex number thermonuclear reaction rate, $\bar{\sigma}$ = complex number thermonuclear reaction cross section which is written as

$$\bar{\sigma} = \sigma \exp(j\theta_\sigma) \quad (328)$$

and where \bar{v} , \bar{p} and $\bar{\epsilon}$ = complex number relative particle speed, relative particle momentum and total particle kinetic energy in the center of mass system respectively. The complex species particle number densities \bar{n}_1 and \bar{n}_2 are written in analogy to equations (61) and (62) as

$$\bar{n}_1 = n_1 \exp(j\theta_{n1}) = \bar{n}_1/V' \quad (329)$$

$$\bar{n}_2 = n_2 \exp(j\theta_{n2}) = \bar{n}_2/V' \quad (330)$$

where

$$n_1 = N_1/V' \cos \theta_{n1} \quad \theta_{n1} = \theta_{n1} \quad (331)$$

$$n_2 = N_2/V' \cos \theta_{n2} \quad \theta_{n2} = \theta_{n2} \quad (332)$$

where θ_{n1} and θ_{n2} = internal phase angles of species particle numbers, N_1 and N_2 = integer number of species particles in the volume V' , and V' = broken symmetry space volume given by equation (6). For the case of coherent spacetime it follows from equation (9) that $V' = V\theta_V$ so that equations (331) and (332) give the species particle number densities as

$$n_1^{\text{coh}} = N_1/(V\theta_V) \cos \theta_{n1}^{\text{coh}} \quad \theta_{n1}^{\text{coh}} = \theta_{n1} \quad (333)$$

$$n_2^{\text{coh}} = N_2/(V\theta_V) \cos \theta_{n2}^{\text{coh}} \quad \theta_{n2}^{\text{coh}} = \theta_{n2} \quad (334)$$

where θ_{n1} and θ_{n2} = internal phase angles for the species particle numbers in coherent spacetime. For incoherent spacetime all internal phase angles are equal to zero and $V' = V$ so that equations (331) and (332) reduce to equation (324). The thermonuclear reaction rates for classical particles, fermions and bosons in the presence of an external gravity field will now be calculated.

A. Thermonuclear Reactions in the Broken Symmetry Spacetime of a Boltzmann Gas in a Gravitational Field.

For a Boltzmann gas in a spacetime with broken internal symmetries due to gravity the complex number thermonuclear reaction rate is given by equation (327)

$$\langle \bar{\sigma} \bar{p} \rangle = \bar{F}_p / \bar{A}_p = F_p / A_p \exp[j(\theta_{Fp} - \theta_{Ap})] \quad (335)$$

$$\langle \bar{\sigma} \bar{\epsilon}^{1/2} \rangle = \bar{F} / \bar{A} = F / A \exp[j(\theta_F - \theta_A)] \quad (336)$$

where

$$\bar{F}_p = \int \bar{\sigma} \bar{p}^3 \exp[-\beta \bar{p}^2 / (2m)] d\bar{p} \quad (337)$$

$$\bar{A}_p = \int \bar{p}^2 \exp[-\beta \bar{p}^2 / (2m)] d\bar{p} \quad (338)$$

$$\bar{F} = \int \bar{\sigma} \bar{\epsilon} \exp(-\beta \bar{\epsilon}) d\bar{\epsilon} \quad (339)$$

$$\bar{A} = \int \bar{\epsilon}^{1/2} \exp(-\beta \bar{\epsilon}) d\bar{\epsilon} \quad (340)$$

where \bar{p} and $\bar{\epsilon}$ are represented by equations (45) and (50) respectively and where $d\bar{p}$ and $d\bar{\epsilon}$ are given by equations (47), (48) and (52), (53) respectively for the general case of partially coherent broken spacetime symmetry and by equation (56) for the case of coherent spacetime. Only the single particle energy integrals appearing in equations (336), (339) and (340) are considered in this paper. A phase space factor FV' given by equations (6) and (77) has been cancelled

from the numerator and denominator of equation (336). The integral \bar{A} that appears in equations (336) and (340) has already appeared in equations (78) and (84) through (89). Using equations (52) and (53) allows the integral \bar{F} that appears in equation (339) to be written as

$$\bar{F} = \int_0^{\infty} \sigma \epsilon \sec \beta_{\epsilon\epsilon} \exp[j(2\theta_{\epsilon} + \theta_{\sigma} + \beta_{\epsilon\epsilon})] \exp(-\beta\bar{\epsilon}) d\epsilon \quad (341)$$

$$= \int_0^{\infty} \sigma \epsilon \sec \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \exp(j\phi_{\sigma\epsilon}) d\epsilon \quad (342)$$

$$= \int_0^{\pi/3} \sigma \epsilon^2 \csc \beta_{\epsilon\epsilon} \exp[j(2\theta_{\epsilon} + \theta_{\sigma} + \beta_{\epsilon\epsilon})] \exp(-\beta\bar{\epsilon}) d\theta_{\epsilon} \quad (343)$$

$$= \int_0^{\pi/3} \sigma \epsilon^2 \csc \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \exp(j\phi_{\sigma\epsilon}) d\theta_{\epsilon} \quad (344)$$

where

$$\phi_{\sigma\epsilon} = 2\theta_{\epsilon} + \theta_{\sigma} - \beta\epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon} \quad (345)$$

The real and imaginary parts of \bar{F} are then obtained from equations (342) and (344) as

$$F_R = \int_0^{\infty} \sigma \epsilon \sec \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \cos \phi_{\sigma\epsilon} d\epsilon \quad (346)$$

$$= \int_0^{\pi/3} \sigma \epsilon^2 \csc \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \cos \phi_{\sigma\epsilon} d\theta_{\epsilon} \quad (347)$$

$$F_I = \int_0^{\infty} \sigma \epsilon \sec \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \sin \phi_{\sigma\epsilon} d\epsilon \quad (348)$$

$$= \int_0^{\pi/3} \sigma \epsilon^2 \csc \beta_{\epsilon\epsilon} \exp(-\beta\epsilon \cos \theta_{\epsilon}) \sin \phi_{\sigma\epsilon} d\theta_{\epsilon} \quad (349)$$

from which F and θ_F are determined to be

$$F = (F_R^2 + F_I^2)^{1/2} \quad \tan \theta_F = F_I / F_R \quad (350)$$

which are required for equation (336). The internal phase angle of the single particle kinetic energy is distributed in the range $0 < \theta_{\epsilon} < \pi/3$ because for a coherent spacetime state with $\beta_{\rho\rho} = \pi/2$ and $\beta_{tt}^0 = \pi/2$ it follows from equations (23) and (51) that $\theta_{\epsilon} = 2(\theta_{\rho} - \theta_t)$ with the upper limits of θ_{ρ} and θ_t being $\theta_{\rho} = \pi/3$ and $\theta_t = \pi/6$.³²

The thermonuclear reaction rate in a Boltzmann gas for the general case of partially coherent spacetime is then obtained from equations (327) and (336) as

$$\begin{aligned}\bar{R} &= \bar{n}_1 \bar{n}_2 (2/m)^{1/2} \bar{F}/\bar{A} \\ &= n_1 n_2 (2/m)^{1/2} F/A \exp[j(\theta_{n1} + \theta_{n2} + \theta_F - \theta_A)]\end{aligned}\quad (351)$$

where A and θ_A are given by equation (79) and F and θ_F are given by equation (350). The real and imaginary parts of equation (351) are given by

$$R_R = n_1 n_2 (2/m)^{1/2} F/A \cos(\theta_{n1} + \theta_{n2} + \theta_F - \theta_A) \quad (352)$$

$$R_I = n_1 n_2 (2/m)^{1/2} F/A \sin(\theta_{n1} + \theta_{n2} + \theta_F - \theta_A) \quad (353)$$

The measured nuclear reaction rate is given by $R_m = R_R$. For the case when all internal phase angles are set equal to zero the measured thermonuclear reaction rate given by equation (352) reduces to the standard incoherent spacetime reaction rate given by equation (325).

In the presence of a very strong gravity field the spacetime can become coherent, and following equation (351) the thermonuclear reaction rate in a Boltzmann gas is written as

$$\bar{R}^{\text{coh}} = \bar{n}_1^{\text{coh}} \bar{n}_2^{\text{coh}} (2/m)^{1/2} \bar{F}^{\text{coh}}/\bar{A}^{\text{coh}} \quad (354)$$

where \bar{A}^{coh} is given by equations (104) through (108), and where for coherent spacetime with $\beta_{\epsilon\epsilon} = \pi/2$ and $\epsilon = \epsilon = \text{constant}$ it follows from equations (343) and (344) that

$$\bar{F}^{\text{coh}} = \epsilon_C^2 \bar{C} = \epsilon_C^2 (C_P + jC_I) = \epsilon_C^2 C \exp(j\theta_C) \quad (355)$$

where

$$\bar{C} = j \int_0^{\pi/3} \sigma \exp[j(2\theta_\epsilon + \theta_\sigma)] \exp(-\beta\bar{\epsilon}) d\theta_\epsilon \quad (356)$$

$$= j \int_0^{\pi/3} \sigma \exp(-\beta\epsilon_C \cos \theta_\epsilon) \exp[j(2\theta_\epsilon + \theta_\sigma - \beta\epsilon_C \sin \theta_\epsilon)] d\theta_\epsilon \quad (357)$$

where $\epsilon_C = \text{constant}$, $\sigma = \sigma(\theta_\epsilon)$ and $\theta_\sigma = \theta_\sigma(\theta_\epsilon)$. The real and imaginary parts of equation (357) are given by

$$C_R = \int_0^{\pi/3} \sigma \exp(-\beta\epsilon_C \cos \theta_\epsilon) \sin(\beta\epsilon_C \sin \theta_\epsilon - \theta_\sigma - 2\theta_\epsilon) d\theta_\epsilon \quad (358)$$

$$C_I = \int_0^{\pi/3} \sigma \exp(-\beta\epsilon_C \cos \theta_\epsilon) \cos(\beta\epsilon_C \sin \theta_\epsilon - \theta_\sigma - 2\theta_\epsilon) d\theta_\epsilon \quad (359)$$

and therefore

$$C = (C_R^2 + C_I^2)^{1/2} \quad \tan \theta_C = C_I/C_R \quad (360)$$

The expression for the thermonuclear reaction rate in a Boltzmann gas that is

located in coherent spacetime is then obtained from equations (105), (354) and (355) as

$$\begin{aligned}\bar{R}^{\text{coh}} &= \bar{n}_1^{\text{coh}} \bar{n}_2^{\text{coh}} (2/m)^{1/2} \epsilon_c^{1/2} \bar{C}/\bar{W} \\ &= n_1^{\text{coh}} n_2^{\text{coh}} (2/m)^{1/2} \epsilon_c^{1/2} C/W \exp[j(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_C - \theta_W)]\end{aligned}\quad (361)$$

where \bar{W} , W and θ_W are given by equations (104) through (110), and \bar{C} , C and θ_C are given by equations (356) through (360). The real and imaginary parts of the thermonuclear reaction rate given in equation (361) are written as

$$R_R^{\text{coh}} = n_1^{\text{coh}} n_2^{\text{coh}} (2/m)^{1/2} \epsilon_c^{1/2} C/W \cos(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_C - \theta_W) \quad (362)$$

$$R_I^{\text{coh}} = n_1^{\text{coh}} n_2^{\text{coh}} (2/m)^{1/2} \epsilon_c^{1/2} C/W \sin(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_C - \theta_W) \quad (363)$$

The measured thermonuclear reaction rate is then given by $R_m^{\text{coh}} = R_R^{\text{coh}}$.

If the fusion cross section is independent of θ_ϵ , it follows from equations (355) through (360) that for coherent spacetime

$$\bar{F}^{\text{coh}} = \bar{\sigma}_c \epsilon_c^2 \bar{D} = \bar{\sigma}_c \epsilon_c^2 (D_R + jD_I) = \sigma_c \epsilon_c^2 D \exp[j(\theta_{\sigma c} + \theta_D)] \quad (364)$$

where the constant fusion cross section is written as

$$\bar{\sigma}_c = \sigma_c \exp(j\theta_{\sigma c}) \quad (365)$$

and where

$$\begin{aligned}\bar{D} &= j \int_0^{\pi/3} \exp(j2\theta_\epsilon) \exp(-\beta\epsilon) d\theta_\epsilon \\ &= j \int_0^{\pi/3} \exp(-\beta\epsilon_c \cos \theta_\epsilon) \exp[j(2\theta_\epsilon - \beta\epsilon_c \sin \theta_\epsilon)] d\theta_\epsilon\end{aligned}\quad (366)$$

The real and imaginary components of equation (366) are given by

$$D_R = \int_0^{\pi/3} \exp(-\beta\epsilon_c \cos \theta_\epsilon) \sin(\beta\epsilon_c \sin \theta_\epsilon - 2\theta_\epsilon) d\theta_\epsilon \quad (367)$$

$$D_I = \int_0^{\pi/3} \exp(-\beta\epsilon_c \cos \theta_\epsilon) \cos(\beta\epsilon_c \sin \theta_\epsilon - 2\theta_\epsilon) d\theta_\epsilon \quad (368)$$

so that

$$D = (D_R^2 + D_I^2)^{1/2} \quad \tan \theta_D = D_I/D_R \quad (369)$$

Then equations (104), (354) and (364) give the thermonuclear reaction rate in a Boltzmann gas that is located in a gravity field that is sufficiently strong as to make spacetime coherent

$$\bar{R}^{\text{coh}} = \bar{n}_1^{\text{coh}} \bar{n}_2^{\text{coh}} (2/m)^{1/2} \bar{\sigma}_c \epsilon_c^{1/2} \bar{D}/\bar{W} \quad (370)$$

$$= \bar{n}_1^{\text{coh}} \bar{n}_2^{\text{coh}} \bar{\sigma}_c v_c \bar{D}/\bar{W} \quad (371)$$

where \bar{D} is given by equation (366), \bar{W} is given by equations (104) and (106), and where v_c = coherent particle velocity which is given by

$$\epsilon_c = 1/2 m v_c^2 = 3/2 k T_c \quad v_c = (3 k T_c / m)^{1/2} \quad (372)$$

where T_c = coherent spacetime temperature for a Boltzmann gas. The quantities $\bar{\sigma}_c$, \bar{D} and \bar{W} are functions of T and T_c

$$\bar{\sigma}_c = \bar{\sigma}_c(T, T_c) \quad \bar{D} = \bar{D}(T, T_c) \quad \bar{W} = \bar{W}(T, T_c) \quad (373)$$

Equation (371) can be rewritten as

$$\bar{R}^{\text{coh}} = \bar{n}_1^{\text{coh}} \bar{n}_2^{\text{coh}} \bar{\sigma}_c (3 k T_c / m)^{1/2} \bar{D}/\bar{W} \quad (374)$$

$$= n_1^{\text{coh}} n_2^{\text{coh}} \sigma_c (3 k T_c / m)^{1/2} D/W \exp[j(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_{\sigma c} + \theta_D - \theta_W)]$$

where W and θ_W are given in equation (110). The real and imaginary parts of the coherent spacetime thermonuclear reaction rate for a Boltzmann gas is obtained from equation (374) as

$$R_R^{\text{coh}} = n_1^{\text{coh}} n_2^{\text{coh}} \sigma_c (3 k T_c / m)^{1/2} D/W \cos(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_{\sigma c} + \theta_D - \theta_W) \quad (375)$$

$$R_I^{\text{coh}} = n_1^{\text{coh}} n_2^{\text{coh}} \sigma_c (3 k T_c / m)^{1/2} D/W \sin(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_{\sigma c} + \theta_D - \theta_W) \quad (376)$$

The measured thermonuclear reaction rate for this case is given by $R_m^{\text{coh}} = R_R^{\text{coh}}$.

The conventional incoherent spacetime calculation of the thermonuclear reaction rate given in equations (325) and (326) should be compared to the coherent spacetime prediction of the thermonuclear reaction rate that is given in equation (375). The coherent spacetime fusion reaction rate is expected to be more sensitive to ambient density and temperature due to the absence of an average of the cross section over a full range of energies that appears in the conventional calculation in equations (325) and (326) but does not appear in equation (375) for the coherent spacetime case. The cross section $\sigma_c(T, T_c)$ is sharply peaked at those temperatures that correspond to resonance energies. Therefore for the coherent spacetime case the energy generation rates in stars will be high only in selected temperature zones that correspond to definite radial distances from the center of a star. Elsewhere in the star the nuclear reaction rates will be low.

B. Thermonuclear Reactions in Fermi Gases in the Presence of Gravity.

For a Fermi gas in broken symmetry spacetime the average quantity that appears in the thermonuclear reaction rate formula given in equation (327) is

evaluated using equation (167) for the complex particle number as follows

$$\langle \bar{\sigma} \bar{\epsilon}^{1/2} \rangle = \bar{K}_N / \bar{K}_D \quad (377)$$

where

$$\bar{K}_N = \int \bar{\sigma} \bar{\epsilon} [\exp(\bar{\epsilon}) + 1]^{-1} d\bar{\epsilon} \quad (378)$$

$$\bar{K}_D = \int \bar{\epsilon}^{1/2} [\exp(\bar{\epsilon}) + 1]^{-1} d\bar{\epsilon} \quad (379)$$

where $\bar{\epsilon}$ is given by equation (168). A factor fV' has been cancelled from the numerator and denominator of equation (377). The series expansion given in equation (169) can be applied to equations (378) and (379) with the result

$$\bar{K}_N = \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}^v \bar{M}_v \quad (380)$$

$$= \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}^v \bar{M}_v \exp[j(v\theta_z + \theta_{Mv})]$$

$$\bar{K}_D = \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}^v \bar{H}_v \quad (381)$$

$$= \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}^v \bar{H}_v \exp[j(v\theta_z + \theta_{Hv})]$$

where \bar{H}_v , H_v and θ_{Hv} are given by equations (171) through (178) and \bar{M}_v is given as follows

$$\bar{M}_v = \int \bar{\sigma} \bar{\epsilon} \exp(-v\beta \bar{\epsilon}) d\bar{\epsilon} = M_v \exp(j\theta_{Mv}) \quad (382)$$

$$= \int_0^{\infty} \sigma \epsilon \sec \beta_{\epsilon\epsilon} \exp(-v\beta \epsilon \cos \theta_{\epsilon}) \exp(j\phi_{\sigma\epsilon v}) d\epsilon \quad (383)$$

$$= \int_0^{\pi/3} \sigma \epsilon^2 \csc \beta_{\epsilon\epsilon} \exp(-v\beta \epsilon \cos \theta_{\epsilon}) \exp(j\phi_{\sigma\epsilon v}) d\theta_{\epsilon} \quad (384)$$

where

$$\phi_{\sigma\epsilon v} = 2\theta_{\epsilon} + \theta_{\sigma} - v\beta \epsilon \sin \theta_{\epsilon} + \beta_{\epsilon\epsilon} \quad (385)$$

The real and imaginary components of equations (383) and (384) are

$$M_{vR} = \int_0^{\infty} \sigma \epsilon \sec \beta_{\epsilon\epsilon} \exp(-v\beta \epsilon \cos \theta_{\epsilon}) \cos(\phi_{\sigma\epsilon v}) d\epsilon \quad (386)$$

$$= \int_0^{\pi/3} \sigma \epsilon^2 \csc \beta_{\epsilon\epsilon} \exp(-v\beta \epsilon \cos \theta_{\epsilon}) \cos(\phi_{\sigma\epsilon v}) d\theta_{\epsilon} \quad (387)$$

$$M_{VI} = \int_0^{\infty} \sigma \epsilon \sec \beta_{\epsilon\epsilon} \exp(-\nu \beta \epsilon \cos \theta_{\epsilon}) \sin(\phi_{\sigma\epsilon\nu}) d\epsilon \quad (388)$$

$$= \int_0^{\pi/3} \sigma \epsilon^2 \csc \beta_{\epsilon\epsilon} \exp(-\nu \beta \epsilon \cos \theta_{\epsilon}) \sin(\phi_{\sigma\epsilon\nu}) d\theta_{\epsilon} \quad (389)$$

from which it follows that

$$M_V = (M_{VR}^2 + M_{VI}^2)^{1/2} \quad \tan \theta_{MV} = M_{VI}/M_{VR} \quad (390)$$

The thermonuclear reaction rate is then given by

$$\bar{R} = \bar{n}_1 \bar{n}_2 (2/m)^{1/2} \bar{K}_N / \bar{K}_D \quad (391)$$

where \bar{K}_N and \bar{K}_D are given by equations (380) and (381). Then the real and imaginary parts of the nuclear reaction rate in a Fermi gas is given by

$$R_R = n_1 n_2 (2/m)^{1/2} K_N / K_D \cos(\theta_{n1} + \theta_{n2} + \theta_{KN} - \theta_{KD}) \quad (392)$$

$$R_I = n_1 n_2 (2/m)^{1/2} K_N / K_D \sin(\theta_{n1} + \theta_{n2} + \theta_{KN} - \theta_{KD}) \quad (393)$$

where

$$K_N = (K_{NR}^2 + K_{NI}^2)^{1/2} \quad \tan \theta_{KN} = K_{NI}/K_{NR} \quad (394)$$

$$K_D = (K_{DR}^2 + K_{DI}^2)^{1/2} \quad \tan \theta_{KD} = K_{DI}/K_{DR} \quad (395)$$

with

$$K_{NR} = \sum_{\nu=1}^{\infty} (-1)^{\nu-1} z^{\nu} M_{\nu} \cos(\nu \theta_z + \theta_{M\nu}) \quad (396)$$

$$K_{NI} = \sum_{\nu=1}^{\infty} (-1)^{\nu-1} z^{\nu} M_{\nu} \sin(\nu \theta_z + \theta_{M\nu}) \quad (397)$$

$$K_{DR} = \sum_{\nu=1}^{\infty} (-1)^{\nu-1} z^{\nu} H_{\nu} \cos(\nu \theta_z + \theta_{H\nu}) \quad (398)$$

$$K_{DI} = \sum_{\nu=1}^{\infty} (-1)^{\nu-1} z^{\nu} H_{\nu} \sin(\nu \theta_z + \theta_{H\nu}) \quad (399)$$

where M_{ν} and $\theta_{M\nu}$ are given by equation (390), and H_{ν} and $\theta_{H\nu}$ are given by equation (178). The measured thermonuclear reaction rate is given by $R_m = R_R$.

Consider now the case of thermonuclear reactions in a Fermi gas that is located in incoherent spacetime (zero gravity field). For this case all internal phase angles are zero and

$$\langle \sigma \epsilon^{1/2} \rangle = K_N^{\text{inc}} / K_D^{\text{inc}} \quad (400)$$

where

$$K_N^{inc} = \sum_{v=1}^{\infty} (-1)^{v-1} z_{inc}^v M_v^{inc} \quad (401)$$

$$K_D^{inc} = \sum_{v=1}^{\infty} (-1)^{v-1} z_{inc}^v H_v^{inc} \quad (402)$$

where H_v^{inc} is given by equation (183) and M_v^{inc} is obtained from equation (386) to be

$$M_v^{inc} = \int_0^{\infty} \sigma \epsilon \exp(-v\beta\epsilon) d\epsilon \quad (403)$$

which is similar to equation (326) for a Boltzmann gas that is located in incoherent spacetime. For the case of incoherent spacetime the thermonuclear reaction rate in a Fermi gas is given by

$$R^{inc} = \eta_1 \eta_2 (2/m)^{1/2} K_N^{inc} / K_D^{inc} \quad (404)$$

where η_1 and η_2 are given by equation (324).

The thermonuclear reaction rate for a Fermi gas that is located in a coherent spacetime state associated with an intense gravity field requires the calculation of the following average value

$$\langle \bar{\sigma} \bar{\epsilon}^{1/2} \rangle = \bar{K}_N^{coh} / \bar{K}_D^{coh} \quad (405)$$

where

$$\begin{aligned} \bar{K}_N^{coh} &= K_N^{coh} \exp(j\theta_{KN}^{coh}) \\ &= j \int_0^{\pi/3} \bar{\sigma} \bar{\epsilon}^2 [\exp(\bar{\xi}_{coh}) + 1]^{-1} d\theta_{\epsilon} \end{aligned} \quad (406)$$

$$\begin{aligned} \bar{K}_D^{coh} &= K_D^{coh} \exp(j\theta_{KD}^{coh}) \\ &= j \int_0^{\pi/3} \bar{\epsilon}^{3/2} [\exp(\bar{\xi}_{coh}) + 1]^{-1} d\theta_{\epsilon} \end{aligned} \quad (407)$$

where

$$\bar{\xi}_{coh} = \beta(\bar{\epsilon} - \bar{\mu}_{coh}) \quad \bar{z}_{coh} = \exp(\beta \bar{\mu}_{coh}) \quad (408)$$

A phase space factor $f v \theta_v$ has been cancelled from the numerator and denominator of equation (405). The integrals in equations (406) and (407) can be rewritten as

$$\begin{aligned} \bar{K}_N^{coh} &= \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{coh}^v \bar{M}_v^{coh} \\ &= \sum_{v=1}^{\infty} (-1)^{v-1} z_{coh}^v M_v^{coh} \exp[j(v\theta_z^{coh} + \theta_{Mv}^{coh})] \end{aligned} \quad (409)$$

$$\begin{aligned}\bar{K}_D^{\text{coh}} &= \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v \bar{H}_v^{\text{coh}} \\ &= \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v H_v^{\text{coh}} \exp[j(v\theta_z^{\text{coh}} + \theta_{Hv}^{\text{coh}})]\end{aligned}\quad (410)$$

where \bar{H}_v^{coh} , H_v^{coh} and θ_{Hv}^{coh} are given by equations (188) through (194), and \bar{M}_v^{coh} is given by

$$\begin{aligned}\bar{M}_v^{\text{coh}} &= M_v^{\text{coh}} \exp(j\theta_{Mv}^{\text{coh}}) \\ &= j\epsilon_c^2 \int_0^{\pi/3} \sigma \exp(-v\beta\epsilon) \exp(j2\theta_\epsilon) d\theta_\epsilon \\ &= j\epsilon_c^2 \int_0^{\pi/3} \sigma \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \exp[j(2\theta_\epsilon + \theta_\sigma - v\beta\epsilon_c \sin \theta_\epsilon)] d\theta_\epsilon\end{aligned}\quad (411)$$

The real and imaginary components of equation (411) are given by

$$M_{vR}^{\text{coh}} = \epsilon_c^2 \int_0^{\pi/3} \sigma \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \sin(v\beta\epsilon_c \sin \theta_\epsilon - 2\theta_\epsilon - \theta_\sigma) d\theta_\epsilon \quad (412)$$

$$M_{vI}^{\text{coh}} = \epsilon_c^2 \int_0^{\pi/3} \sigma \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \cos(v\beta\epsilon_c \sin \theta_\epsilon - 2\theta_\epsilon - \theta_\sigma) d\theta_\epsilon \quad (413)$$

and therefore

$$M_v^{\text{coh}} = [(M_{vR}^{\text{coh}})^2 + (M_{vI}^{\text{coh}})^2]^{1/2} \quad (414)$$

$$\tan \theta_{Mv}^{\text{coh}} = M_{vI}^{\text{coh}} / M_{vR}^{\text{coh}} \quad (415)$$

The component forms of equation (409) are given by

$$K_{NR}^{\text{coh}} = \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v M_v^{\text{coh}} \cos(v\theta_z^{\text{coh}} + \theta_{Mv}^{\text{coh}}) \quad (416)$$

$$K_{NI}^{\text{coh}} = \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v M_v^{\text{coh}} \sin(v\theta_z^{\text{coh}} + \theta_{Mv}^{\text{coh}}) \quad (417)$$

and the component forms of equation (410) are

$$K_{DR}^{\text{coh}} = \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v H_v^{\text{coh}} \cos(v\theta_z^{\text{coh}} + \theta_{Hv}^{\text{coh}}) \quad (418)$$

$$K_{DI}^{\text{coh}} = \sum_{v=1}^{\infty} (-1)^{v-1} z_{\text{coh}}^v H_v^{\text{coh}} \sin(v\theta_z^{\text{coh}} + \theta_{Hv}^{\text{coh}}) \quad (419)$$

which gives

$$K_N^{\text{coh}} = [(K_{NR}^{\text{coh}})^2 + (K_{NI}^{\text{coh}})^2]^{1/2} \quad (420)$$

$$K_D^{\text{coh}} = [(K_{DR}^{\text{coh}})^2 + (K_{DI}^{\text{coh}})^2]^{1/2} \quad (421)$$

$$\tan \theta_{KN}^{\text{coh}} = K_{NI}^{\text{coh}} / K_{NR}^{\text{coh}} \quad (422)$$

$$\tan \theta_{KD}^{\text{coh}} = K_{DI}^{\text{coh}} / K_{DR}^{\text{coh}} \quad (423)$$

Then the general expression for the thermonuclear reaction rate for a Fermi gas in coherent spacetime is obtained from equations (327) and (405) through (423) to be

$$\bar{R}^{\text{coh}} = \bar{n}_1^{\text{coh}} \bar{n}_2^{\text{coh}} (2/m)^{1/2} \bar{K}_N^{\text{coh}} / \bar{K}_D^{\text{coh}} \quad (424)$$

whose real and imaginary parts are written as

$$R_R^{\text{coh}} = n_1^{\text{coh}} n_2^{\text{coh}} (2/m)^{1/2} K_N^{\text{coh}} / K_D^{\text{coh}} \cos(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_{KN}^{\text{coh}} - \theta_{KD}^{\text{coh}}) \quad (425)$$

$$R_I^{\text{coh}} = n_1^{\text{coh}} n_2^{\text{coh}} (2/m)^{1/2} K_N^{\text{coh}} / K_D^{\text{coh}} \sin(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_{KN}^{\text{coh}} - \theta_{KD}^{\text{coh}}) \quad (426)$$

The measured reaction rate is given for this case by $R_m^{\text{coh}} = R_R^{\text{coh}}$.

Some simplification can be made for the case of fusion reactions occurring in a Fermi gas with coherent spacetime if the thermonuclear cross section is a constant which is independent of the internal phase angles, and for this case equations (411) through (415) become

$$\bar{M}_v^{\text{coh}} = \bar{\sigma}_c \epsilon_c^2 \bar{D}_v \quad (427)$$

$$M_{vR}^{\text{coh}} = \sigma_c \epsilon_c^2 D_v \cos(\theta_{\sigma c} + \theta_{Dv}) \quad (428)$$

$$M_{vI}^{\text{coh}} = \sigma_c \epsilon_c^2 D_v \sin(\theta_{\sigma c} + \theta_{Dv}) \quad (429)$$

$$M_v^{\text{coh}} = \sigma_c \epsilon_c^2 D_v \quad (430)$$

$$\theta_{Mv}^{\text{coh}} = \theta_{\sigma c} + \theta_{Dv} \quad (431)$$

where

$$\begin{aligned} \bar{D}_v &= j \int_0^{\pi/3} \exp(-v\beta\bar{\epsilon}) \exp(j2\theta_\epsilon) d\theta_\epsilon \\ &= j \int_0^{\pi/3} \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \exp[j(2\theta_\epsilon - v\beta\epsilon_c \sin \theta_\epsilon)] d\theta_\epsilon \end{aligned} \quad (432)$$

The real and imaginary parts of equation (432) are given by

$$D_{vR} = \int_0^{\pi/3} \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \sin(v\beta\epsilon_c \sin \theta_\epsilon - 2\theta_\epsilon) d\theta_\epsilon \quad (433)$$

$$D_{vI} = \int_0^{\pi/3} \exp(-v\beta\epsilon_c \cos \theta_\epsilon) \cos(v\beta\epsilon_c \sin \theta_\epsilon - 2\theta_\epsilon) d\theta_\epsilon \quad (434)$$

from which it follows that

$$D_v = (D_{vR}^2 + D_{vI}^2)^{1/2} \quad \tan \theta_{Dv} = D_{vI}/D_{vR} \quad (435)$$

Therefore for this case equations (409) and (410) can be written for the Fermi gas in coherent spacetime as

$$\bar{K}_N^{\text{coh}} = \bar{\sigma}_c \epsilon_c^2 \bar{G}_N^{\text{coh}} \quad (436)$$

$$\bar{K}_D^{\text{coh}} = \epsilon_c^{3/2} \bar{G}_D^{\text{coh}} \quad (437)$$

where

$$\bar{G}_N^{\text{coh}} = G_N^{\text{coh}} \exp(j\theta_{GN}^{\text{coh}}) = \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v \bar{D}_v \quad (438)$$

$$\bar{G}_D^{\text{coh}} = G_D^{\text{coh}} \exp(j\theta_{GD}^{\text{coh}}) = \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v \bar{E}_v \quad (439)$$

where equations (437) and (439) are obtained from equation (410) by using equation (188) where \bar{E}_v is given by equation (190). The magnitudes and phase angles that appear in equations (438) and (439) are calculated as follows

$$G_N^{\text{coh}} = [(G_{NR}^{\text{coh}})^2 + (G_{NI}^{\text{coh}})^2]^{1/2} \quad (440)$$

$$G_D^{\text{coh}} = [(G_{DR}^{\text{coh}})^2 + (G_{DI}^{\text{coh}})^2]^{1/2} \quad (441)$$

$$\tan \theta_{GN}^{\text{coh}} = G_{NI}^{\text{coh}}/G_{NR}^{\text{coh}} \quad (442)$$

$$\tan \theta_{GD}^{\text{coh}} = G_{DI}^{\text{coh}}/G_{DR}^{\text{coh}} \quad (443)$$

where from equations (438) and (439)

$$G_{NR}^{\text{coh}} = \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v D_v \cos(v\theta_z^{\text{coh}} + \theta_{Dv}) \quad (444)$$

$$G_{NI}^{\text{coh}} = \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v D_v \sin(v\theta_z^{\text{coh}} + \theta_{Dv}) \quad (445)$$

$$G_{DR}^{\text{coh}} = \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v E_v \cos(v\theta_z^{\text{coh}} + \theta_{Ev}) \quad (446)$$

$$G_{DI}^{\text{coh}} = \sum_{v=1}^{\infty} (-1)^{v-1} \bar{z}_{\text{coh}}^v E_v \sin(v\theta_z^{\text{coh}} + \theta_{Ev}) \quad (447)$$

where D_v and θ_{Dv} are given by equation (435), and E_v and θ_{Ev} are given by equation (194). The thermonuclear reaction rate for a Fermi gas in coherent space-

time is then obtained from equations (424), (436) and (437) as

$$\begin{aligned}
 \bar{R}^{\text{coh}} &= \bar{n}_1^{\text{coh}} \bar{n}_2^{\text{coh}} (2/m)^{1/2} \bar{K}_N^{\text{coh}} / \bar{K}_D^{\text{coh}} \\
 &= \bar{n}_1^{\text{coh}} \bar{n}_2^{\text{coh}} (2/m)^{1/2} \bar{\sigma}_c \epsilon_c^{1/2} \bar{G}_N^{\text{coh}} / \bar{G}_D^{\text{coh}} \\
 &= \bar{n}_1^{\text{coh}} \bar{n}_2^{\text{coh}} v_c \bar{\sigma}_c \bar{G}_N^{\text{coh}} / \bar{G}_D^{\text{coh}} \\
 &= n_1^{\text{coh}} n_2^{\text{coh}} v_c \sigma_c G_N^{\text{coh}} / G_D^{\text{coh}} \exp[j(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_{\sigma c} + \theta_{GN}^{\text{coh}} - \theta_{GD}^{\text{coh}})]
 \end{aligned} \tag{448}$$

where n_1^{coh} , n_2^{coh} , θ_{n1}^{coh} and θ_{n2}^{coh} are given by equations (333) and (334), and v_c = coherent spacetime relative particle speed given by equation (372). The real and imaginary parts of the complex number thermonuclear reaction rate given by equation (448) are

$$R_R^{\text{coh}} = n_1^{\text{coh}} n_2^{\text{coh}} v_c \sigma_c G_N^{\text{coh}} / G_D^{\text{coh}} \cos(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_{\sigma c} + \theta_{GN}^{\text{coh}} - \theta_{GD}^{\text{coh}}) \tag{449}$$

$$R_I^{\text{coh}} = n_1^{\text{coh}} n_2^{\text{coh}} v_c \sigma_c G_N^{\text{coh}} / G_D^{\text{coh}} \sin(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_{\sigma c} + \theta_{GN}^{\text{coh}} - \theta_{GD}^{\text{coh}}) \tag{450}$$

The measured thermonuclear reaction rate is given by $R_m^{\text{coh}} = R_R^{\text{coh}}$.

C. Thermonuclear Reactions in a Bose Gas Located in a Gravitational Field.

For the case of thermonuclear reactions in a Bose gas situated in space-time with gravity induced broken internal coordinate symmetries the reaction rate formula in equation (327) requires an averaged quantity which is evaluated using the complex number Bose-Einstein distribution as follows

$$\langle \bar{\sigma} \bar{\epsilon}^{1/2} \rangle = \bar{L}_N / \bar{L}_D \tag{451}$$

where

$$\bar{L}_N = L_N \exp(j\theta_{LN}) \tag{452}$$

$$= \sum_p \bar{\sigma}_p \bar{\epsilon}_p^{1/2} [1/\bar{z} \exp(\beta \bar{\epsilon}_p) - 1]^{-1} (fV')^{-1} \tag{453}$$

$$= \int \bar{\sigma} \bar{\epsilon} [\exp(\bar{\epsilon}) - 1]^{-1} d\bar{\epsilon} \tag{454}$$

$$\bar{L}_D = L_D \exp(j\theta_{LD}) \tag{455}$$

$$= \sum_p [1/\bar{z} \exp(\beta \bar{\epsilon}_p) - 1]^{-1} (fV')^{-1} \tag{456}$$

$$= \int \bar{\epsilon}^{1/2} [\exp(\bar{\epsilon}) - 1]^{-1} d\bar{\epsilon} + \bar{z}(1 - z)^{-1} (fV')^{-1} \tag{457}$$

where $\bar{\epsilon}$ and f are given by equations (168) and (77) respectively. The reciprocal

square bracket terms in equations (452) through (457) can be expanded in a power series in a manner similar to that in equations (284) through (286) with the result that

$$\bar{L}_N = \sum_{v=1}^{\infty} \bar{z}^v \bar{M}_v \quad (458)$$

$$\bar{L}_D = \sum_{v=1}^{\infty} \bar{z}^v \bar{H}_v + \bar{z}(1 - \bar{z})^{-1} (fV')^{-1} \quad (459)$$

where \bar{M}_v is given by equation (382) and \bar{H}_v is given by equation (171). From equations (452) through (459) it follows that

$$L_N = (L_{NR}^2 + L_{NI}^2)^{1/2} \quad (460)$$

$$L_D = (L_{DR}^2 + L_{DI}^2)^{1/2} \quad (461)$$

$$\tan \theta_{LN} = L_{NI}/L_{NR} \quad (462)$$

$$\tan \theta_{LD} = L_{DI}/L_{DR} \quad (463)$$

where

$$L_{NR} = \sum_{v=1}^{\infty} z^v M_v \cos(v\theta_z + \theta_{Mv}) \quad (464)$$

$$L_{NI} = \sum_{v=1}^{\infty} z^v M_v \sin(v\theta_z + \theta_{Mv}) \quad (465)$$

$$L_{DR} = \sum_{v=1}^{\infty} z^v H_v \cos(v\theta_z + \theta_{Hv}) + g \quad (466)$$

$$L_{DI} = \sum_{v=1}^{\infty} z^v H_v \sin(v\theta_z + \theta_{Hv}) + h \quad (467)$$

with

$$g = (fV')^{-1} z(\cos \theta_z - z)/(1 + z^2 - 2z \cos \theta_z) \quad (468)$$

$$h = (fV')^{-1} (z \sin \theta_z)/(1 + z^2 - 2z \cos \theta_z) \quad (469)$$

In the limit of high temperatures equations (71) and (72) show that $z \approx 1$ and $\theta_z \approx 0$ and therefore $g \approx 0$ and $h \approx 0$. Then the thermonuclear reaction rate for a Bose gas in the presence of gravity is obtained from equations (327) and (451) to be

$$\bar{R} = \bar{n}_1 \bar{n}_2 (2/m)^{1/2} \bar{L}_N / \bar{L}_D \quad (470)$$

The real and imaginary parts of the thermonuclear reaction rate are given by

$$R_R = n_1 n_2 (2/m)^{1/2} L_N/L_D \cos(\theta_{n1} + \theta_{n2} + \theta_{LN} - \theta_{LD}) \quad (471)$$

$$R_I = n_1 n_2 (2/m)^{1/2} L_N/L_D \sin(\theta_{n1} + \theta_{n2} + \theta_{LN} - \theta_{LD}) \quad (472)$$

The measured reaction rate is given by $R_m = R_R$.

For incoherent spacetime the thermonuclear reaction rate of a Bose gas is obtained by first noting that equation (451) becomes

$$\langle \sigma \epsilon^{1/2} \rangle_{\text{inc}} = L_N^{\text{inc}} / L_D^{\text{inc}} \quad (473)$$

where

$$L_N^{\text{inc}} = \sum_{v=1}^{\infty} z_{\text{inc}}^v M_v^{\text{inc}} \quad (474)$$

$$L_D^{\text{inc}} = \sum_{v=1}^{\infty} z_{\text{inc}}^v H_v^{\text{inc}} + (fV')^{-1} z_{\text{inc}} (1 - z_{\text{inc}})^{-1} \quad (475)$$

where M_v^{inc} and H_v^{inc} are given by equations (403) and (183). Finally, the nuclear reaction rate for this case is obtained from equations (327) and (473) as

$$R^{\text{inc}} = n_1 n_2 (2/m)^{1/2} L_N^{\text{inc}} / L_D^{\text{inc}} \quad (476)$$

which is the standard result.

For thermonuclear reactions in a Bose gas in the presence of gravity that is so strong as to make the spacetime coherent, the relevant averaged quantity that appears in equation (327) is written as

$$\langle \bar{\sigma} \bar{\epsilon}^{1/2} \rangle = \bar{L}_N^{\text{coh}} / \bar{L}_D^{\text{coh}} = \bar{L}_N^{\text{coh}} / \bar{L}_D^{\text{coh}} \exp[j(\theta_{LN}^{\text{coh}} - \theta_{LD}^{\text{coh}})] \quad (477)$$

where

$$\bar{L}_N^{\text{coh}} = j \int_0^{\pi/3} \bar{\sigma} \bar{\epsilon}^2 [\exp(\bar{\xi}_{\text{coh}}) - 1]^{-1} d\theta_{\epsilon} \quad (478)$$

$$\bar{L}_D^{\text{coh}} = j \int_0^{\pi/3} \bar{\epsilon}^{3/2} [\exp(\bar{\xi}_{\text{coh}}) - 1]^{-1} d\theta_{\epsilon} + (fV\theta_V)^{-1} \bar{z}_{\text{coh}} (1 - \bar{z}_{\text{coh}})^{-1} \quad (479)$$

where $\bar{\xi}_{\text{coh}}$ is given by equation (408). Expanding the integrands in a power series yields

$$\bar{L}_N^{\text{coh}} = \sum_{v=1}^{\infty} \bar{z}_{\text{coh}}^v \bar{M}_v^{\text{coh}} \quad (480)$$

$$\bar{L}_D^{\text{coh}} = \sum_{v=1}^{\infty} \bar{z}_{\text{coh}}^v \bar{H}_v^{\text{coh}} + (fV\theta_V)^{-1} \bar{z}_{\text{coh}} (1 - \bar{z}_{\text{coh}})^{-1} \quad (481)$$

The component forms of equations (480) and (481) are written as

$$L_{NR}^{coh} = \sum_{v=1}^{\infty} z_{coh}^v M_v^{coh} \cos(v\theta_z^{coh} + \theta_{Mv}^{coh}) \quad (482)$$

$$L_{NI}^{coh} = \sum_{v=1}^{\infty} z_{coh}^v M_v^{coh} \sin(v\theta_z^{coh} + \theta_{Mv}^{coh}) \quad (483)$$

$$L_{DR}^{coh} = \sum_{v=1}^{\infty} z_{coh}^v H_v^{coh} \cos(v\theta_z^{coh} + \theta_{Hv}^{coh}) + g_{coh} \quad (484)$$

$$L_{DI}^{coh} = \sum_{v=1}^{\infty} z_{coh}^v H_v^{coh} \sin(v\theta_z^{coh} + \theta_{Hv}^{coh}) + h_{coh} \quad (485)$$

where \bar{M}_v^{coh} , M_v^{coh} and θ_{Mv}^{coh} are given by equations (411), (414) and (415); \bar{H}_v^{coh} , H_v^{coh} and θ_{Hv}^{coh} are given by equations (188) and (189); z_{coh} and θ_z^{coh} are given by equations (197) and (198), and where in analogy to equations (468) and (469)

$$g_{coh} = (fv\theta_v)^{-1} z_{coh} (\cos \theta_z^{coh} - z_{coh}) / (1 + z_{coh}^2 - 2z_{coh} \cos \theta_z^{coh}) \quad (486)$$

$$h_{coh} = (fv\theta_v)^{-1} (z_{coh} \sin \theta_z^{coh}) / (1 + z_{coh}^2 - 2z_{coh} \cos \theta_z^{coh}) \quad (487)$$

The magnitudes and phase angles that appear in equations (477) are given by

$$L_N^{coh} = [(L_{NR}^{coh})^2 + (L_{NI}^{coh})^2]^{1/2} \quad (488)$$

$$L_D^{coh} = [(L_{DR}^{coh})^2 + (L_{DI}^{coh})^2]^{1/2} \quad (489)$$

$$\tan \theta_{LN}^{coh} = L_{NI}^{coh} / L_{NR}^{coh} \quad (490)$$

$$\tan \theta_{LD}^{coh} = L_{DI}^{coh} / L_{DR}^{coh} \quad (491)$$

Then the general expression for the thermonuclear reaction rate in a Bose gas located in coherent spacetime is given by

$$\bar{R}^{coh} = \bar{n}_1^{coh} \bar{n}_2^{coh} (2/m)^{1/2} \bar{L}_N^{coh} / \bar{L}_D^{coh} \quad (492)$$

$$R_R^{coh} = n_1^{coh} n_2^{coh} (2/m)^{1/2} L_N^{coh} / L_D^{coh} \cos(\theta_{n1}^{coh} + \theta_{n2}^{coh} + \theta_{LN}^{coh} - \theta_{LD}^{coh}) \quad (493)$$

$$R_I^{coh} = n_1^{coh} n_2^{coh} (2/m)^{1/2} L_N^{coh} / L_D^{coh} \sin(\theta_{n1}^{coh} + \theta_{n2}^{coh} + \theta_{LN}^{coh} - \theta_{LD}^{coh}) \quad (494)$$

The measured reaction rate is given by $R_m^{coh} = R_R^{coh}$.

For the case when the thermonuclear reaction cross section for a Bose gas

in coherent spacetime is a constant which is independent of θ_c , then equations (427), (480) and (481) give

$$\bar{L}_N^{\text{coh}} = \bar{\sigma}_c \bar{\epsilon}_c^2 \bar{C}_N^{\text{coh}} \quad (495)$$

$$\bar{L}_D^{\text{coh}} = \epsilon_c^{3/2} \bar{C}_D^{\text{coh}} + (fv\theta_v)^{-1} \bar{z}_{\text{coh}} (1 - \bar{z}_{\text{coh}})^{-1} \quad (496)$$

where

$$\begin{aligned} \bar{C}_N^{\text{coh}} &= C_N^{\text{coh}} \exp(j\theta_{CN}^{\text{coh}}) = \sum_{v=1}^{\infty} \bar{z}_{\text{coh}}^v \bar{D}_v \\ &= \sum_{v=1}^{\infty} z_{\text{coh}}^v D_v \exp[j(v\theta_z^{\text{coh}} + \theta_{Dv})] \end{aligned} \quad (497)$$

$$\begin{aligned} \bar{C}_D^{\text{coh}} &= C_D^{\text{coh}} \exp(j\theta_{CD}^{\text{coh}}) = \sum_{v=1}^{\infty} \bar{z}_{\text{coh}}^v \bar{E}_v \\ &= \sum_{v=1}^{\infty} z_{\text{coh}}^v E_v \exp[j(v\theta_z^{\text{coh}} + \theta_{Ev})] \end{aligned} \quad (498)$$

where \bar{D}_v and \bar{E}_v are given by equations (432) and (190). From equations (497) and (498) it follows that

$$C_{NR}^{\text{coh}} = \sum_{v=1}^{\infty} z_{\text{coh}}^v D_v \cos(v\theta_z^{\text{coh}} + \theta_{Dv}) \quad (499)$$

$$C_{NI}^{\text{coh}} = \sum_{v=1}^{\infty} z_{\text{coh}}^v D_v \sin(v\theta_z^{\text{coh}} + \theta_{Dv}) \quad (500)$$

$$C_{DR}^{\text{coh}} = \sum_{v=1}^{\infty} z_{\text{coh}}^v E_v \cos(v\theta_z^{\text{coh}} + \theta_{Ev}) \quad (501)$$

$$C_{DI}^{\text{coh}} = \sum_{v=1}^{\infty} z_{\text{coh}}^v E_v \sin(v\theta_z^{\text{coh}} + \theta_{Ev}) \quad (502)$$

and therefore

$$C_N^{\text{coh}} = [(C_{NR}^{\text{coh}})^2 + (C_{NI}^{\text{coh}})^2]^{1/2} \quad (503)$$

$$C_D^{\text{coh}} = [(C_{DR}^{\text{coh}})^2 + (C_{DI}^{\text{coh}})^2]^{1/2} \quad (504)$$

$$\tan \theta_{CN}^{\text{coh}} = C_{NI}^{\text{coh}} / C_{NR}^{\text{coh}} \quad \tan \theta_{CD}^{\text{coh}} = C_{DI}^{\text{coh}} / C_{DR}^{\text{coh}} \quad (505)$$

For the case of a constant fusion cross section $\bar{\sigma}_c$ independent of θ_c it follows from equations (495) and (496) that for a Bose gas in coherent spacetime

$$L_{NR}^{coh} = \sigma_c \epsilon_c^2 C_N^{coh} \cos(\theta_{\sigma c} + \theta_{CN}^{coh}) \quad (506)$$

$$L_{NI}^{coh} = \sigma_c \epsilon_c^2 C_N^{coh} \sin(\theta_{\sigma c} + \theta_{CN}^{coh}) \quad (507)$$

$$L_{DR}^{coh} = \epsilon_c^{3/2} C_D^{coh} \cos \theta_{CD}^{coh} + g_{coh} \quad (508)$$

$$L_{DI}^{coh} = \epsilon_c^{3/2} C_D^{coh} \sin \theta_{CD}^{coh} + h_{coh} \quad (509)$$

where g_{coh} and h_{coh} are given by equations (486) and (487). From equations (488), (506) and (507) and from equations (489), (508) and (509) it follows that

$$L_N^{coh} = \sigma_c \epsilon_c^2 C_N^{coh} \quad (510)$$

$$L_D^{coh} = [\epsilon_c^3 (C_D^{coh})^2 + 2\epsilon_c^{3/2} C_D^{coh} (g_{coh} \cos \theta_{CD}^{coh} + h_{coh} \sin \theta_{CD}^{coh}) + g_{coh}^2 + h_{coh}^2]^{1/2} \quad (511)$$

Combining equations (490), (506) and (507) gives

$$\theta_{LN}^{coh} = \theta_{\sigma c} + \theta_{CN}^{coh} \quad (512)$$

while θ_{LD}^{coh} is obtained from equations (491), (508) and (509)

$$\tan \theta_{LD}^{coh} = (\epsilon_c^{3/2} C_D^{coh} \sin \theta_{CD}^{coh} + h_{coh}) / (\epsilon_c^{3/2} C_D^{coh} \cos \theta_{CD}^{coh} + g_{coh}) \quad (513)$$

Then equations (492) through (494) coupled with equations (510) through (513) give the thermonuclear reaction rate for a Bose gas in coherent spacetime and for a constant fusion cross section.

If the ground state of the Bose gas in coherent spacetime is unoccupied due to high temperatures then $z_{coh} = 1$ and $\theta_z^{coh} = 0$ from equations (71) and (72), so that equations (486) and (487) give $g_{coh} = 0$ and $h_{coh} = 0$ and equation (511) becomes

$$L_D^{coh} = \epsilon_c^{3/2} C_D^{coh} \quad (514)$$

while equation (513) gives for this case

$$\theta_{LD}^{coh} = \theta_{CD}^{coh} \quad (515)$$

Combining equations (493), (494), (510), (512), (514) and (515) gives the real and imaginary parts of the thermonuclear reaction rate as

$$R_R^{\text{coh}} = n_1^{\text{coh}} n_2^{\text{coh}} v_c \sigma_c C_N^{\text{coh}} / C_D^{\text{coh}} \cos(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_{\sigma c} + \theta_{CN}^{\text{coh}} - \theta_{CD}^{\text{coh}}) \quad (516)$$

$$R_I^{\text{coh}} = n_1^{\text{coh}} n_2^{\text{coh}} v_c \sigma_c C_N^{\text{coh}} / C_D^{\text{coh}} \sin(\theta_{n1}^{\text{coh}} + \theta_{n2}^{\text{coh}} + \theta_{\sigma c} + \theta_{CN}^{\text{coh}} - \theta_{CD}^{\text{coh}}) \quad (517)$$

where C_N^{coh} , θ_{CN}^{coh} , C_D^{coh} and θ_{CD}^{coh} are defined in equations (497) through (505), and where v_c is given by equation (372).

There are three reasons for the lowered rates of some thermonuclear reactions compared to the standard reaction rate calculations. The first reason is that gravity induces a broken spacetime symmetry, and the thermonuclear reaction cross section for coherent spacetime is evaluated at the prevailing single value of the kinetic energy per nucleus ϵ_c so that, unless this kinetic energy is associated with a resonance value for the cross section, the thermonuclear reaction rate will be low compared to the incoherent spacetime calculation of the standard model which assumes a distribution of kinetic energies. The second reason for the lowered thermonuclear reaction rate is the appearance of a cosine factor in the predicted measured values of the reaction rate. The argument of the cosine term consists of the sum of the internal phase angles of the two reacting species particle number densities, the internal phase angle of the fusion cross section, and the internal phase angles associated with the coherent spacetime average of the relative speed of the interacting particles. The third reason for the depressed thermonuclear reaction rate compared to the standard prediction is due to the effects of a real gas state equation which enters the renormalization group time equation given in equations (1A) and (1B).

6. CONCLUSION. It is suggested that in very strong gravitational fields spacetime becomes coherent and that the motion in such a state occurs internally so that the space and time coordinates, single particle momenta, and single particle energies change coherently in that they rotate in internal space with their magnitudes held fixed. The state equations for the noninteracting Boltzmann, Fermi and Bose gases are developed for partially and totally coherent spacetime. These complex number state equations reduce to the standard ideal gas state equations for the case of incoherent spacetime. Thermonuclear reaction rates in bulk matter are lowered, compared to the standard predictions which assume an ideal gas, by the real gas nature of the state equations of stellar matter as described by a renormalization group time equation. These reaction rates will also be lowered by the degree of coherence of spacetime because the values of the integrals over momentum or kinetic energy that are required to calculate the thermonuclear reaction rates are dependent on the state of coherence of spacetime. The thermonuclear reaction rates have been evaluated for the ideal Boltzmann gas, Fermi gas and Bose gas for partially coherent and totally coherent spacetime. The lowered thermonuclear reaction rates are due to real gas effects in the relativistic equation for time and to the effects of the coherence of spacetime on the reaction rate integrals for the ideal gases, and may represent an explanation for the depressed values of the measured solar neutrino emission flux compared to predictions of the standard solar model.

ACKNOWLEDGEMENT

Special thanks go to Elizabeth K. Klein for typing and editing this paper.

REFERENCES

1. Rolfs, C. E. and Rodney, W. S., Cauldrons in the Cosmos, Univ. of Chicago Press, Chicago, 1988.
2. Aller, L. H., Astrophysics, Vols. 1 & 2, Ronald Press, New York, 1954.
3. Stein, R. F., and Cameron, A. G. W., Stellar Evolution, Plenum, New York, 1966.
4. Shapiro, S. L. and Teukolsky, S. A., Highlights of Modern Astrophysics, John Wiley, New York, 1986.
5. Fowler, W. A. and Hoyle, F., Nucleosynthesis in Massive Stars and Supernovae, Univ. of Chicago Press, Chicago, 1964.
6. Schwarzschild, M., Structure and Evolution of the Stars, Dover, New York, 1958.
7. Van Horn, H. M., "Dense Astrophysical Plasmas," *Science*, Vol. 252, p. 384, 19 April 1991.
8. Trimble, V., "The Origin and Abundances of the Chemical Elements," *Revs. Mod. Phys.*, Vol. 47, p. 877, October 1975.
9. Woosley, S. E. and Weaver, T. A., "The Physics of Supernova Explosions," *Ann. Rev. Astron. Astrophys.*, Vol. 24, p. 205, 1986.
10. Fowler, W. A., "Experimental and Theoretical Nuclear Astrophysics: The Quest for the Origin of the Elements," *Rev. Mod. Phys.*, Vol. 56, p. 149, April 1984.
11. Woosley, S. E. and Phillips, M. M., "Supernova 1987A," *Science*, Vol. 240, p. 750, 6 May 1988.
12. Trimble, V., "1987A: The Greatest Supernova Since Kepler," *Rev. Mod. Phys.*, Vol. 60, p. 859, October 1988.
13. Arnett, W. D., Bahcall, J. N., Kirshner, R. P. and Woosley, S. E., "Supernova 1987A," *Ann. Rev. Astron. Astrophys.*, Vol. 27, p. 629, 1989.
14. Boesgaard, A. M. and Steigman, G., "Big Bang Nucleosynthesis: Theories and Observations," *Ann. Rev. Astron. Astrophys.*, Vol. 23, p. 319, 1985.
15. Kolb, E. W., Turner, M. S., Chakravorty, A. and Schramm, D. N., "Constraints from Primordial Nucleosynthesis on the Mass of the τ Neutrino," *Phys. Rev. Lett.*, Vol. 67, p. 533, 29 July 1991.
16. Bahcall, J. N. and Ulrich, R. K., "Solar Models, Neutrino Experiments, and Helioseismology," *Rev. Mod. Phys.*, Vol. 60, p. 297, April 1988.

17. Dearborn, D. S. and Fuller, G. M., "Neutrino Oscillations and Uncertainty in the Solar Model," Phys. Rev. D, Vol. 39, p. 3543, 15 June 1989.
18. Wolfenstein, L. and Beier, E., "Neutrino Oscillations and Solar Neutrinos," Physics Today, p. 28, July 1989.
19. Bethe, H. A., "Solar-Neutrino Experiments," Phys. Rev. Lett., Vol. 63, 21 August 1989.
20. Hirata, K. S., et al., "Observation of ^8B Solar Neutrinos in the Kamio-kande-II Detector," Phys. Rev. Lett., Vol. 63, 3 July 1989.
21. Fukugita, M. and Yanagida, T., "Possible Solution to the Discrepancy between the Homestake and Kamiokande Solar-Neutrino Experiments," Phys. Rev. Lett., Vol. 65, p. 1975, 15 October 1990.
22. Bahcall, J. N. and Bethe, H. A., "A Solution of the Solar-Neutrino Problem," Phys. Rev. Lett., Vol. 65, p. 2233, 29 October 1990.
23. Hirata, K. S., et al., "Results from One Thousand Days of Real-Time, Directional Solar-Neutrino Data," Phys. Rev. Lett., Vol. 65, p. 1297, 10 September 1990.
24. Abazov, A. I., et al., "Search for Neutrinos from the Sun Using the Reaction $^{71}\text{Ga}(\nu_e, e^-)^{71}\text{Ge}$," Phys. Rev. Lett., Vol. 67, p. 3332, 9 December 1991.
25. Minakata, H. and Nunokawa, H., "Hybrid Solution of the Solar Neutrino Problem in Anticorrelation with Sunspot Activity," Phys. Rev. Lett., Vol. 63, p. 121, 10 July 1989.
26. Babu, K. S. and Mohapatra, R. N., "Model for Large Transition Magnetic Moment of the Electron Neutrino," Phys. Rev. Lett., Vol. 63, p. 228, 17 July 1989.
27. Bahcall, J. N. and Haxton, J. N., "Matter-Enhanced Neutrino Oscillations in the Standard Model," Phys. Rev. D, Vol. 40, p. 931, 15 August 1989.
28. Babu, K. S. and Mohapatra, R. N., "Geometrical Neutrino Mass Hierarchy and a 17-keV ν_τ ," Phys. Rev. Lett., Vol. 67, p. 1498, 16 September 1991.
29. Bethe, H. A. and Bahcall, J. N., "Solar Neutrinos and the Mikheyev-Smirnov-Wolfenstein Theory," Phys. Rev. D, Vol. 44, p. 2962, 15 November 1991.
30. Weiss, R. A., "Gauge Theory of Time," Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 91-1, p. 367, June 19-22, 1990.
31. Weiss, R. A., "Quantum Theory of Time and Thermodynamics," Ninth Army Conference on Applied Mathematics and Computing, Univ. of Minnesota, Minneapolis, MN, ARO 92-1, June 18-21, 1991, p. 565.

32. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.
33. Huang, K., Statistical Mechanics, John Wiley, New York, 1963.
34. Mayer, J. E. and Mayer, M. G., Statistical Mechanics, John Wiley, New York, 1977.
35. Hill, T. L., Statistical Thermodynamics, Addison-Wesley, Reading, MA, 1960.
36. Weiss, R. A., Relativistic Thermodynamics, Exposition Press, New York, 1976.
37. Gradshteyn, I. S. and Ryzhik, I. M., Table of Integrals, Series, and Products, Academic Press, New York, 1980.

Some methods of analysis in the study of microstructure

David Kinderlehrer

Center for Nonlinear Analysis and Department of Mathematics
Carnegie Mellon University
Pittsburgh, PA 15213-3890

1. Introduction Fine scale morphology or microstructure is implicated in the macroscopic behavior of many materials, but the manifestations of this are often unclear¹. We are in need of improved methods for studying this frequently encountered situation. In this report we describe in an expository fashion the initial developments of one such technique which has been applied in several instances especially related to certain alloys or other crystalline materials. Good examples where defect structures consisting of fine scale morphology are relatively simple are certain phase transformations of displacive or structural type and the mechanical behavior of shape memory alloys. Martensitic materials, in particular, exhibit fine twinned microstructures, often appearing as layers or layers within layers². Although we often refer to microstructure, we are confronted with a primarily continuum phenomenon for which some authors use the term mesoscale. In these considerations, one issue is paramount: the presence of spatially oscillatory behavior and the means of understanding it constitutes the bridge from the fine scale to the large scale.

Crystals are idealized as materials with a high degree of configurational order. As a consequence, the continuum energy densities ascribed to them are invariant under discrete groups and have multiple potential wells. Such densities are not lower semicontinuous. The infimum of energy may be obtained only in some generalized sense, while a minimizing sequence may develop successively finer oscillations. Said in another fashion, when the material deforms owing to change in its environment, the configurational order acts as a constraint resulting in the creation of a defect structure, which in this case is a complicated spatially oscillatory fine structure. The limit deformation alone need not be sufficient to characterize many of the properties of the limit configuration.

¹ Supported in part by the Army Research Office.

² For illustrations of oscillatory behavior in alloys and other materials see [1,2,3,4,5,27,53].

A feature of the constitutive theory under discussion is that surface energies, magnetic domain wall energies, and similar effects are neglected, although the highly nonlinear potential well structure for the material has a prominent role. Thus fine phase laminar twin systems and fine phase magnetic domain structures may tend to limits of infinite fineness. The theory in this formulation delivers useful information about variant arrangement and location as well as macroscopic state functions like energy and stress. It is particularly useful in deciding where in the body fine structure will arise.

At the analytical level, we apply a recently developed averaging method, briefly explained in §2 below, which accounts for rapidly spatially varying systems and accomodates the fine scale microstructure. A configuration which minimizes a given variational principle is described in terms of generalized moments of the minimizing sequence, or equivalently, oscillatory statistics. The most important property of the method is to unify energetic and kinematic considerations by compelling the statistics to be consistent with the variational principle.

Examples of this sort of analysis served to generalize the crystallographic theory of martensite, Ball and James [1], and to compute the relaxation of energy densities in the presence of symmetry, Chipot and Kinderlehrer [9] and Fonseca [26]. It has subsequently played a role in many discussions related to microstructure, eg. [2,4,5,10,12,13,14,15,25,27,28,30,31,32,33,35,39,40,41,45,47]. A treatment of the variational foundations of this method is given in [29,36,37,38]. Kohn [42,43] has shown how these ideas and those of relaxation in, general are consistent with the treatment of Khachaturyan and Roitburd, eg. [34,49]. Here we shall briefly explore two examples: a theory for highly magnetostrictive iron/rare earth alloys and a mathematical example of evolution of fine structure. A major impetus for these investigations is to provide a basis for the numerical computation of configurations with complicated microstructure. A few selected results of these efforts will be reported.

2. Local spatial averages Young measures. We describe the portrayal of microstructure or fine structure by local spatial averages or Young measures . We also explain the mechanism by which these averages serve to unify energetic and kinematic considerations. Since this may not be familiar to most readers, we give some examples as well. A bounded sequence of functions or more general fields, scalar, vector, or matrix valued,

$$f^k: \Omega \rightarrow \mathbb{R}^N, k = 1,2,\dots, \quad (2.1)$$

may describe a spatially oscillatory structure or system in the region Ω . For example, if Ω is a cube, f_0 a fixed periodic function, and

$$f^k(x) = f_0(kx),$$

the system represents spatial oscillations modulated in some fashion by f_0 . A specific one dimensional example is

$$f^k(x) = \sin \pi kx, \quad 0 \leq x \leq 1, \quad k = 1, 2, 3, \dots \quad (2.2)$$

Another one is

$$f^k(x) = \begin{cases} -1 & \frac{j-1}{k} < x < \frac{j}{k} \\ 1 & \frac{j}{k} \leq x \leq \frac{j+1}{k} \end{cases} \quad 1 \leq j \leq k, \quad 0 \leq x \leq 1. \quad (2.3)$$

The general sequence (f^k) may fail to converge pointwise or even in the mean, as the examples (2.2) and (2.4) above illustrate. This, it turns out, is characteristic of the minimizing sequences for functionals which lack lower semicontinuity and in particular of variational problems associated to crystalline solids in the context of finite elasticity.

The behavior of the sequence may be grasped by computing limits of averages

$$\bar{f}(a) = \lim_{\rho \rightarrow \infty} \lim_{k \rightarrow \infty} \frac{1}{|B_\rho|} \int_{B_\rho(a)} f^k dx, \quad (2.4)$$

where $|B_\rho|$ stands for the volume of the ball of radius ρ . This tells us only the average limit of the sequence, however, and does not inform us of its particular oscillatory behavior. The technical name for this convergence is *weak* convergence*. To overcome this, we calculate generalized moments. Let ψ be any continuous function and consider the sequence $(\psi(f^k))$. Although this sequence need not converge, we may ascertain, as above, a weak limit function

$$\bar{\psi}(a) = \lim_{\rho \rightarrow \infty} \lim_{k \rightarrow \infty} \frac{1}{|B_\rho|} \int_{B_\rho(a)} \psi(f^k) dx. \quad (2.5)$$

The association

$$\psi \rightarrow \bar{\psi}(a)$$

gives rise to an integral representation (a probability measure) on ψ ,

$$\bar{\psi}(a) = \int_{\mathbb{R}^N} \psi(\lambda) dv_a(\lambda) \quad (2.6)$$

which has the property

$$\int_E \psi(f^k) dx \rightarrow \int_E \bar{\psi} dx \quad \text{for any subset } E \subset \Omega. \quad (2.7)$$

This collection of measures $v = (v_x)_{x \in \Omega}$ summarizes the statistics of the spatial oscillations of the sequence. It was introduced by Young [54] to study control problems. Its first use in differential equations is due to Tartar [50,51] who studied hyperbolic conservation laws. They are measures defined on the range of the sequence (f^k) which depend on the point $x \in \Omega$.

In particular, it is generally incorrect to suppose that the limit of a minimizing sequence realizes the infimum of energy in a variational principle whose minimizing sequences are highly oscillatory. The minimum energy must be evaluated using (2.6).

Examples

For example, both the sequences of (2.2) and (2.3) have $\bar{f}(x) = 0$. On the other hand, for (2.2),

$$\bar{\psi}(a) = \frac{1}{\pi} \int_{-1}^1 \psi(\lambda) \frac{d\lambda}{\sqrt{1-\lambda^2}}, \quad 0 < a < 1, \quad (2.8)$$

while for (2.3),

$$\bar{\psi}(a) = \frac{1}{2} (\psi(-1) + \psi(1)), \quad 0 < a < 1. \quad (2.9)$$

The oscillatory statistics of the two sequences are thus quite different.

Let us now give a simple well known example of how oscillations may arise in the mathematical context. The first of these is the familiar Young-Zermelo tacking problem, [54]. Let $\phi(\lambda)$ be a double well potential with equal wells at -1 and 1 as depicted in Figure 1 and, with $\Omega = (0,1)$ an interval, set

$$I(v) = \int_{\Omega} (\varphi(v') + v^2) dx. \quad (2.10)$$

A minimizing sequence (u^k) for this functional wishes to enjoy both $\frac{du^k}{dx} = \pm 1$ for all k and $u^k \rightarrow 0$ in Ω . The result is the generation of oscillations, with a typical minimizing sequence given by u^k with

$$\frac{du^k}{dx} = f^k \quad \text{in } \Omega, \quad (2.11)$$

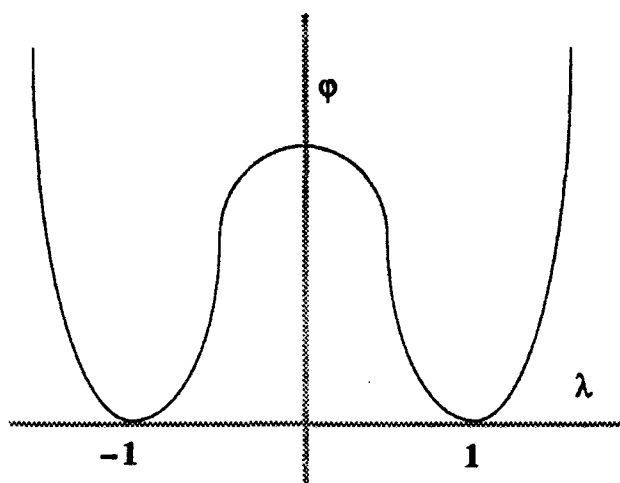


Figure 1 A typical double well potential in one variable.

with the f^k defined in (2.3). The *Young measure solution* of the minimization problem is given by, cf. (2.9),

$$v = \frac{1}{2}(\delta_{-1} + \delta_{+1}). \quad (2.12)$$

In this example, oscillations are created by competition between the two terms of the functional. In multivariable problems, side conditions, like boundary conditions, are sufficient to give rise to an oscillatory structure.

Interestingly, it is difficult to decide this from a computational standpoint because the additional.

competition between the grid orientation and the particular kinematics organization of the configuration requires a sufficiently large computational domain as well as certain other features. We are examining these issues with Nicolaides.

The propagation of oscillations, and even the convection of oscillations is an important issue. Tartar has investigated this in some detail [52], recently introducing the H measure to account for aspects of the frequency distribution of a sequence as well.

3. Magnetostriction A remarkable feature of ferromagnetic materials is that the single domain state is generally unstable. This contrasts with martensite, where the single variant configuration is stable for arbitrarily large samples. In the blue phase of cholesteric liquid crystals, the failure of stability of the uniform state relative to an array of defects is termed *frustration*. Our theory here could be interpreted as one possible interpretation of this phenomenon at a macroscopic

scale. The frustration in our system arises from the competition of an anisotropy energy which demands constant magnetization strength and direction with an induced field energy which prefers to tend to zero. A consequence of this is to promote development of a fine scale structure which seeks to compromise the constraint of constant magnetization strength.

Certain iron/rare earth alloys display both frustration and a huge magnetostriction. There are cubic Laves phase RFe_2 (R = rare earth) compounds, for example, where magnetically induced strains "overwhelm the conventional thermal expansion of the material", Clark [11]. TbDyFe_2 (terfenol) solidifies from the melt with a complex highly mobile domains consisting of structural domains and magnetic domains. Typical growth habits result in configurations with parallel twinned layers, cf. Figure 2, that persist in the magnetostrictive process. We have been studying this with a theory of magnetoelastic interactions based on the micromagnetics of W. F. Brown, Jr. [6,7,8] and the symmetry considerations introduced by Ericksen [16-24]. For a complete discussion, we refer to James and Kinderlehrer [32]. It has some similarities with Toupin's theory of the elastic dielectric [54]. We then apply it to the equilibrium microstructure of TbDyFe_2 . The primary mechanism of magnetostriction appears to be an exchange of stability of mechanical variants under the influence of a change in the magnetic field, but we do not discuss this in detail here.

For relatively rigid materials one may assume the free energy to depend on magnetization alone, [30,31]. The theory in this case gives good qualitative agreement with experiment, explaining why cubic magnets have a few large domains and why uniaxial ones have a fine structure. Domain refinement at the boundary is also predicted when the normal to the boundary has a suitable orientation with relative to the crystal axes, in agreement with observations.

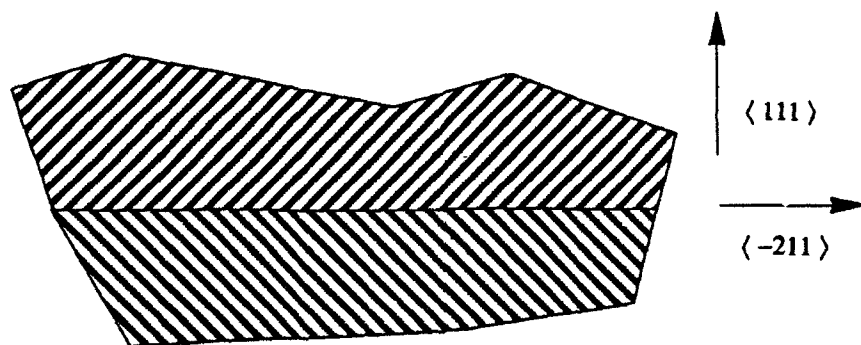


Figure 2. Schematic depiction of the microstructure in a sample of TbDyFe_2 illustrating the herringbone structure of two sets of laminar fine structures. Crystallographic directions are with reference to the high temperature nonmagnetic phase.

The variational principle is formulated in terms of a stored energy density which depends on the deformation gradient $F \in M$, 3×3 matrices, magnetization (per unit mass) $m \in \mathbb{R}^3$, and temperature $\theta \in \mathbb{R}$. We suppose it given by a nonnegative function

$$W(F, m, \theta) \quad F \in M, m \in \mathbb{R}^3, \theta \in \mathbb{R}, \quad (3.1)$$

subject to the condition of frame indifference

$$W(QF, mQ^T, \theta) = W(F, m, \theta), \quad Q \in SO(3), \quad (3.2)$$

and material symmetry

$$W(FP, m, \theta) = W(F, m, \theta), \quad P \in \mathcal{P}, \quad (3.3)$$

where \mathcal{P} is a crystallographic point group.

Requiring W to depend on the deformation gradient $F = \nabla y$ and magnetization m but not on $\nabla^2 y$ and ∇m indicates that any energy associated with mechanical twin walls and Bloch walls is neglected. In this formulation, there may be infinitely fine twins or infinitely fine magnetic domains, as we have suggested earlier. Since on a macroscopic level, the materials of interest display highly mobile domain configurations, any wall energies need be very small. The analytical benefit is that in the limit of infinite fineness we are able to determine rather accurately the arrangement and location of variants within the material, although not their dimensions.

Let y denote the spatial variable and H and M denote the magnetic field and the magnetization (dipole moment per unit volume), respectively. In the spatial configuration, Maxwell's equations hold. In addition, material is magnetically saturated. For an appropriate choice of units, and introducing $U(y)$ as a potential for H ,

$$\operatorname{div}_y (-\nabla_y U + M) = 0 \quad \text{in } \mathbb{R}^3, \quad (3.4)$$

and the field energy density is given by

$$\frac{1}{2} |H|^2 = \frac{1}{2} |\nabla_y U|^2.$$

The saturation constraint leads to

$$\left| \frac{M}{\rho} \right| = f(\theta) \quad \text{in the body}, \quad (3.5)$$

where ρ is the density.

The domain Ω is interpreted as an undistorted single crystal above the Curie temperature. By an abuse of notation, let $y(x)$ denote the deformation of Ω to $y(\Omega)$, assumed for the purposes of discussion to be 1:1. Since $\rho(x) = 1/\det \nabla y(x)$, the magnetization per unit mass previously introduced, $m = \det \nabla y M$, so the constraint (3.5) assumes the form

$$|m| = f(\theta).$$

We assume $f(\theta) = 1$, without loss in generality.

In this fashion we may write the virtual energy of the configuration $y = y(x)$, $m = m(x)$ in the mixed reference/spatial form

$$E(y,m) = \int_{\Omega} W(\nabla y, m, \theta) dx + \frac{1}{2} \int_{\mathbb{R}^3} |\nabla_y U|^2 dy \quad (3.6)$$

subject to the constraints,

$$\operatorname{div}_y (-\nabla_y U + \frac{1}{\det \nabla y} m) = 0 \quad \text{in } \mathbb{R}^3. \quad (3.7)$$

$$|m| = 1 \quad \text{in } y(\Omega).$$

From (3.7), we may also write the energy in the form

$$E(y,m) = \int_{\Omega} W(\nabla y, m, \theta) dx + \frac{1}{2} \int_{y(\Omega)} \frac{1}{\det \nabla y} m \cdot \nabla_y U dy. \quad (3.8)$$

Both for computational and analytical reasons, it is useful to express this in terms of reference variables alone. For this, introduce $u(x) = U(y(x))$, so $\nabla u(x) = \nabla_y U(y(x))F(x)$, $F(x) = \nabla y(x)$. With $C = F^T F$, the constraint equation (2.9) becomes

$$\operatorname{div}(-\nabla u C^{-1} \det F + m F^{-T}) = 0 \quad \text{in } \mathbb{R}^3, \quad (3.9)$$

and the saturation condition is simply

$$|m| = 1 \quad \text{in } \Omega. \quad (3.10)$$

The virtual energy of $y = y(x)$, $m = m(x)$ in reference form is

$$E(y,m) = \int_{\Omega} W(\nabla y, m, \theta) dx + \frac{1}{2} \int_{\mathbb{R}^3} \nabla u \cdot C^{-1} \cdot \nabla u \det F dx, \quad (3.11)$$

subject to (3.9) and (3.10). Analogous to (3.8), we may also write (3.11) as

$$E(y,m) = \int_{\Omega} W(\nabla y, m, \theta) dx + \frac{1}{2} \int_{\mathbb{R}^3} \nabla u \cdot m F^{-T} dx. \quad (3.12)$$

The symmetry condition (3.3) induces a potential well structure on W . The arrangement of these potential wells determines the possible fine structure. Our schema for understanding this well structure begins by choosing for \mathcal{P} the symmetry group of a putative high temperature non-magnetic parent phase of the material. For example, in the case we shall consider here, \mathcal{P} is the cubic group of order 24: relative to a cubic basis, these are the proper orthogonal matrices of the form $P = (p_{ij})$, $p_{ij} = \pm 1$ or 0. This is the appropriate assumption for TbDyFe_2 . For $\theta < \theta_0$, we assume there exists a pair (U_1, m_1) with $|m_1| = 1$ and $U_1 = U_1^T$ positive definite satisfying

$$W(U_1, m_1, \theta) \leq W(F, m, \theta) \quad \text{for } F \in D, |m| = 1. \quad (3.13)$$

Generally, U_1 and m_1 depend on temperature. The conditions (3.2) and (3.3) imply the existence of other minima by (2.9). Assume that *the full set of minima is determined by the orbits of (U_1, m_1) under these actions*. Thus

$$\begin{aligned} \inf W &= W(RU_1 H, m_1 R^T, \theta) < W(F, m, \theta) \quad \text{for } R \in \text{SO}(3), H \in \mathcal{P} \\ &\text{and } F \in M, |m| = 1, \text{ with } (F, m) \neq (RU_1 H, m_1 R^T). \end{aligned} \quad (3.14)$$

The potential wells are described as

$$(RU_1, m_1 R^T), \quad R \in \text{SO}(3),$$

$$(RU_2, m_2 R^T), \quad R \in \text{SO}(3),$$

$$(RU_n, m_n R^T), \quad R \in \text{SO}(3),$$

where

$$\{(U_1, m_1), (U_2, m_2), \dots (U_n, m_n)\} = \{(QU_1 Q^T, m_1 Q^T) : Q \in \mathcal{P}\}.$$

An orbit of the form $(RU_i, m_i R^T)$, $R \in SO(3)$, will be called a *variant* by analogy to martensitic transformations.

Our idea of a variational principle is to find a pair (y, m) such that

$$E(y, m) = \inf \{ E(\eta, \mu) : (\eta, \mu) \text{ subject to (3.9)} \}.$$

However, in our situation, with the material, in essence, uniaxial, this will not be possible. Instead we must content ourselves with this result, whose verification relies on an explicit construction:

$$\inf E = \min W | \Omega |. \quad (3.15)$$

4. The variational context

4.1 The variational context: energetics

Consider the minimization question associated to (3.8) subject to (3.9). By choosing a special sequence of magnetizations, one may show that

$$\inf E(y, m) = \min W | \Omega |, \quad (4.1)$$

as discussed at the end of §3. However, because of the competition between the field energy and the stored energy, there cannot be any pair (y^*, m^*) with y^* affine and

$$E(y^*, m^*) = \min W | \Omega |. \quad (4.2)$$

We are led in this manner to consider a sequence of deformation fields and magnetization fields (y^k, m^k) subject to (3.9) for which (dependence on θ suppressed)

$$E(y^k, m^k) \rightarrow \min W | \Omega |. \quad (4.3)$$

and

$$\nabla y^k \rightarrow \nabla \bar{y} \quad \text{and} \quad m^k \rightarrow \bar{m},$$

where the convergence is in the sense of (2.4), or equivalently, (2.7).

The only way for (4.3) to occur is if

$$W(y^k, m^k) \rightarrow \min W \quad \text{and} \quad \frac{1}{2} \int_{\mathbb{R}^3} |\nabla_y U^k|^2 dy \rightarrow 0. \quad (4.4)$$

Since

$$W(y^k, m^k) \rightarrow \bar{W}(x), \quad \text{for } x \in \Omega,$$

$$\bar{W}(x) = \int_{M \times S^2} W(A, \mu) dv_x(A, \mu),$$

we must have that the set of (A, μ) charged by v , that is the support of the measure v , is contained in the minimum energy wells described by (3.17). In analytical terms we write

$$\text{supp } v \subset \{(A, \mu): W(A, \mu) = \min W\} = \Sigma. \quad (4.5)$$

In addition, (4.4) provides via the constraint equation in (3.9) that

$$\text{div}_y \frac{1}{\det \nabla y^k} m^k \rightarrow 0 \quad \text{in } H^{-1}(\mathbb{R}^3). \quad (4.6)$$

(4.5) and (4.6) place severe constraints on the possible forms of $\nabla \bar{y}$ and \bar{m} .

4.2 The variational context: kinematics

An easy integration by parts shows that if (y^k) is a sequence of deformation fields with bounded derivatives, then for any minors $M(\nabla y^k)$ of the matrices ∇y^k ,

$$M(\nabla y^k) \rightarrow M(\nabla \bar{y})$$

in the sense of (3.4), that is, in the weak* sense. Thus minors are special functions $\psi(A)$ which are continuous with respect to this convergence. They are, of course, the null-Lagrangians. The Young measure relation also holds. So, in the present situation, combining (4.5) with the Young measure representation gives

$$\nabla \bar{y}(x) = \int_{\Sigma} A dv_x(A, \mu), \quad (4.7)$$

$$\text{adj } \nabla \bar{y}(x) = \int_{\Sigma} \text{adj } A dv_x(A, \mu), \quad \text{and} \quad (4.8)$$

$$\det \nabla \bar{y}(x) = \int_{\Sigma} \det A \, dv_x(A, \mu), \quad (4.9)$$

where $\text{adj } A$ stands for the classical adjoint of A and $\det A$ stands for the determinant of A . Formula (4.7) is simply a restatement of (3.4) in this case and is included to provide a complete list of null-lagrangians. We refer to (4.7) - (4.9) as the *minors relations*.

These relations place extremely strong restrictions on the nature of possible equilibrium configurations because they assert that the limit statistics of equilibrium configurations must be compatible with the potential well structure of the macroscopic bulk energy.

It is worthwhile pointing out that for the special case of an infinitely fine laminate supported on two deformation gradients M_1 and M_2 , that is,

$$\int_{\Sigma} \psi(A) \, dv_x(A, \mu) = (1 - \theta) \psi(M_1) + \theta \psi(M_2), \quad (4.10)$$

for some θ , $0 < \theta < 1$, the minors relations imply that

$$M_2 - M_1 = a \otimes n = \text{rank one}, \quad (4.11)$$

and the $\{M_i\}$ may represent the deformation gradients of twin related variants with normal n . A sequence of deformations which determines (4.10) with $\theta = \frac{1}{2}$ is given by

$$\nabla y^k(x) = M_1 + \frac{1}{2}(1 + f^k(x \cdot n))a \otimes n, \quad x \in \Omega, \quad (4.12)$$

where $f^k(t)$ is defined in (2.3).

Analogous formulas hold for any problem in thermoelasticity, but in magnetostriction we also have a relation about magnetization owing to (3.12). This relation is most useful in reference coordinates. Recall that

$$\bar{m}(x) = \int_{\Sigma} \mu \, dv_x(A, \mu). \quad (4.13)$$

The Phase Transition in TbDyFe₂

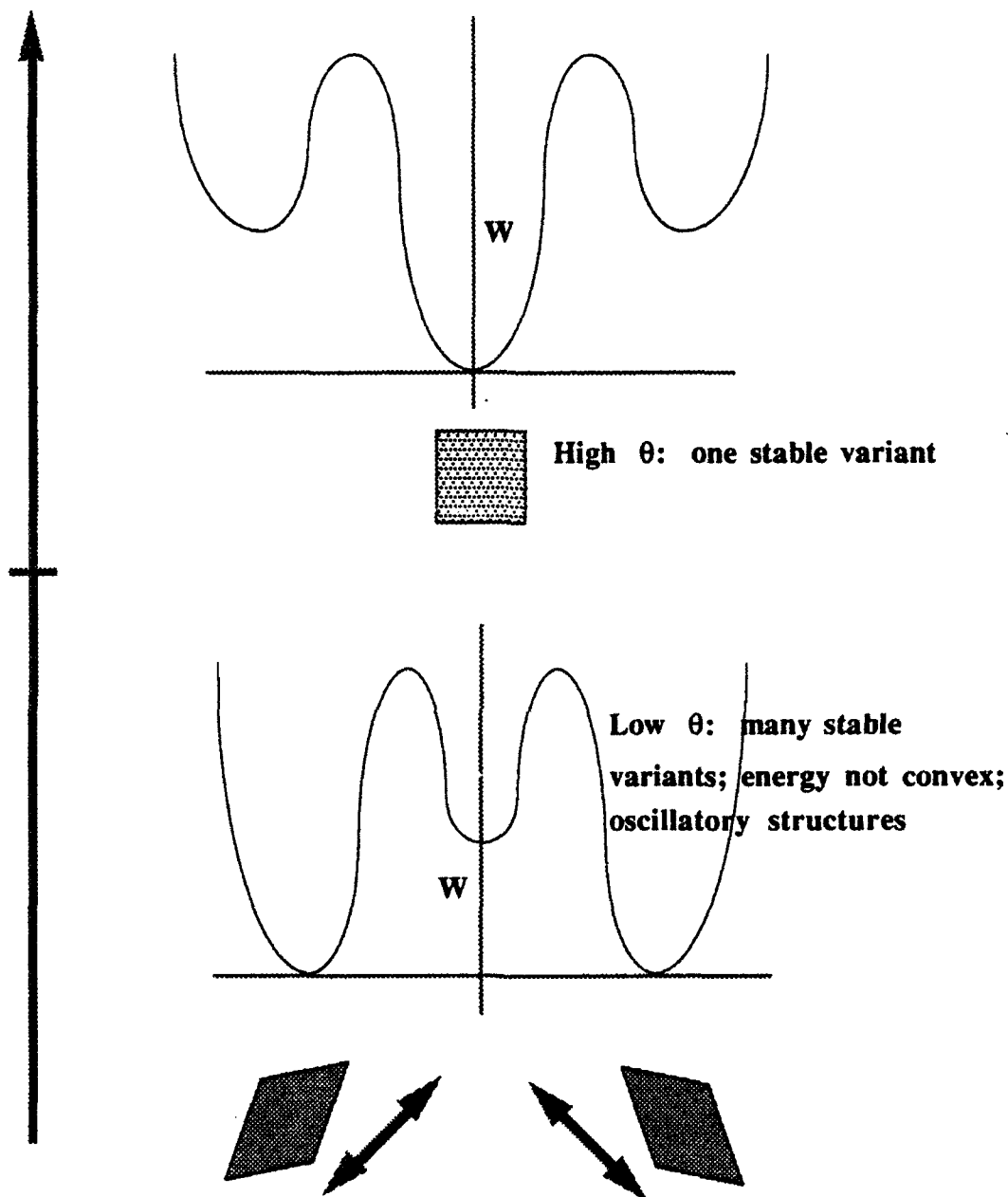


Figure 3 The phase transition in terfenol.

The new relation is that

$$\bar{m}(x)\nabla\bar{y}(x)^{-T} = \int_{\Sigma} \mu A^{-T} dv_x(A,\mu), \quad (4.14)$$

with

$$\operatorname{div}(\bar{m}(\nabla \bar{y})^{-T}) = 0 \quad \text{in } \mathbb{R}^3. \quad (4.15)$$

4.3 Application

In Terfenol-D, onset of ferromagnetism is associated with a stretch of the unit cell along a main diagonal parallel to the magnetization. In simplified kinematics, we find a pair (U_1, m_1) satisfying (2.15) provided by

$$U_1 = 1 + \varepsilon m_1 \otimes m_1 \quad \text{and} \quad m_1 = \frac{1}{\sqrt{3}}(1, 1, 1), \quad |\varepsilon| \text{ small}, \quad (4.1)$$

for a suitable choice of coordinates. The other potential wells are determined by

$$U_i = 1 + \varepsilon m_i \otimes m_i, \quad i = 2, 3, 4, \quad \text{with} \quad (4.2)$$

$$m_2 = \frac{1}{\sqrt{3}}(-1, 1, 1), \quad m_3 = \frac{1}{\sqrt{3}}(1, -1, 1), \quad m_4 = \frac{1}{\sqrt{3}}(1, 1, -1).$$

Since $-m_j$ is also an admissible magnetization, there are a total of eight potential wells. We regard the coordinates chosen so that this represents the lower laminate in Figure 1. The upper laminate is obtained from it by a rotation about the m_1 axis. Note that this is not a symmetry operation of the original energy and, although holding invariant the well of (U_1, m_1) , gives a different set of wells. To save space, here we discuss only the lower laminate. To properly treat the entire system, we must introduce an inhomogeneous energy $W(F, m, \theta, x)$, $x \in \Omega$, cf. [32].

To establish our result we wish to check that we may produce a minimizing sequence (y^k, m^k) for the energy $E(y, m)$ with the potential well structure determined by (4.1) and (4.2) whose statistics, as determined by the "minors relations" and their generalizations, (3.13) - (3.15), (3.20), and (3.21), deliver the observed crystallographic data, for example, of the lower laminate of Figure 1. We are able to do this using the wells determined by (U_1, m_1) and (U_2, m_2) .

Given any pair of transformation strains

$$U_1 = 1 + \varepsilon \xi_1 \otimes \xi_1 \quad \text{and} \quad U_2 = 1 + \varepsilon \xi_2 \otimes \xi_2,$$

$$|\xi_i| = 1, \quad \xi_1 \text{ and } \xi_2 \text{ independent,}$$

then the type I and type II twins (or twins and reciprocal twins) have normals

$$n^+ = \xi_1 + \xi_2 \text{ and } n^- = \xi_1 - \xi_2.$$

There are rotations $R^\pm(\epsilon)$ and vectors $a^\pm(\epsilon)$ with

$$U_1 = R^\pm U_2 (1 + a^\pm \otimes n^\pm) \quad (4.3)$$

In this case with $\xi_i = m_i$, $n^+ = \langle 100 \rangle$ and $n^- = \langle 011 \rangle$, in agreement with the observations of D. Lord [44,53].

A coherent laminate may be constructed from the deformation gradients U_1 and R^+U_2 or from the deformation gradients U_1 and R^-U_2 , cf. also (3.16) - (3.19). We may construct a compatible sequence of magnetizations m^k with $m^k = \pm m_1$ in the U_1 regions and $m^k = \pm m_2(R^+)^T$ in the R^+U_2 regions with the property that the limit average $\bar{m} = 0$ so that

$$\lim E(y^k, m^k) = \min W |\Omega|,$$

cf. Figure 2 below.

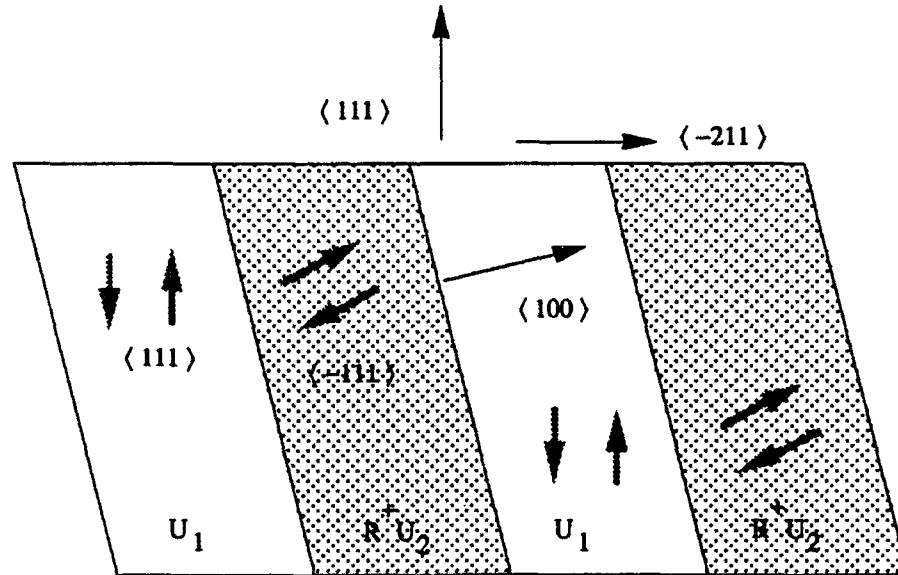


Figure 4. The equilibrium microstructure of a laminate with parameters predicted by the theory. The gray arrows represent directions of the magnetization within the mechanical layers. In the U_1 layers they are $\pm m_1$ where m_1 is a $\langle 111 \rangle$ direction and in the shaded layers they are $\pm m_2(R^+)^T$ where m_2 is a $\langle -111 \rangle$ direction.

It is possible to deduce, moreover, that the only magnetization distributions consistent with the mechanical laminate have $\bar{m} = 0$. One may explicitly write a Young measure solution

$$v_x = \frac{1}{2}(1-\lambda)(\delta_{(U_1, m_1)} + \delta_{(U_1, -m_1)}) + \frac{1}{2}\lambda(\delta_{(R+U_2, m_2(R)^T)} + \delta_{(R+U_2, -m_2(R)^T)}), x \in \Omega^-,$$

where $0 \leq \lambda \leq 1$.

Our analysis suggests however that Figure 2 above is not the only solution and need not be the only one the laboratory photographs show either. A laminate may also be realized with deformation gradients U_3 and RU_4 which has the same appearance on an $\langle 01-1 \rangle$ plane. Note that $m_3 + m_4 \parallel (100)$. This configuration has the property that it is exactly compatible across the $\langle 111 \rangle$ plane whereas compatibility of the U_1 and RU_2 laminate is only in the fine structure limit and requires λ constant. Interestingly, the fine structure laminate might display greater magnetostriction.

The computation of configurations is underway by Ma, who has successfully reproduced hysteretic behavior in linearly magnetostrictive models. Previously, Luskin and Ma, [45], studied the rigid ferromagnet.

5. Evolution Evolution problems for potentials which are not convex may be considered within this framework. The basics of an existence theory were given in [40] and has been significantly advanced by Demoulini and Walkington, in work which has not been published. Walkington, in particular, has adapted method for computing solutions of the Young-Zermelo problem to evolution, for reasons which will become clear momentarily.

Consider φ as in Figure 1, a scalar valued potential for example, and ask for a solution of the problem, $q(\lambda) = \varphi_\lambda(\lambda)$,

$$\begin{aligned} -\operatorname{div} q(\nabla u) + \frac{\partial u}{\partial t} &= 0 \quad \text{in } \Omega \times (0, \infty) \\ u|_{\Omega \times (0)} &= u_0 \\ u|_{\partial\Omega \times (0, \infty)} &= 0 \end{aligned} \tag{5.1}$$

A classical solution need not exist because the equation may be backward parabolic in some regions, but we may seek a Young measure solutions along the lines we have been discussing. We find this solution by adapting an implicit scheme.

The functional

$$I(v) = \int_{\Omega} (\varphi(\nabla v) + \frac{1}{2h}(v - w)^2) dx, \quad h > 0, \quad (5.2)$$

is only slightly different from (2.10). Given $h > 0$, set $t_k = kh$ and $u^{h,0} = u_0$. Solve iteratively for Young measures $v^{h,k}$ and underlying deformations $u^{h,k}$ by the minimization procedure

$$\int_{\Omega} (\varphi(\nabla v) + \frac{1}{2h}(v - u^{h,k-1})^2) dx \rightarrow \min \quad (5.3)$$

where the competing $v \in H_0^1(\Omega)$, for example. The $v^{h,k}$ and $u^{h,k}$ satisfy

$$\int_{\Omega} (\bar{\varphi} + \frac{1}{2h}(v - u^{h,k-1})^2) dx = \min, \quad (5.4)$$

$$\bar{\varphi}(x) = \int_{\mathbb{R}^N} \varphi(\lambda) dv_x^{h,k}(\lambda), \quad \text{with}$$

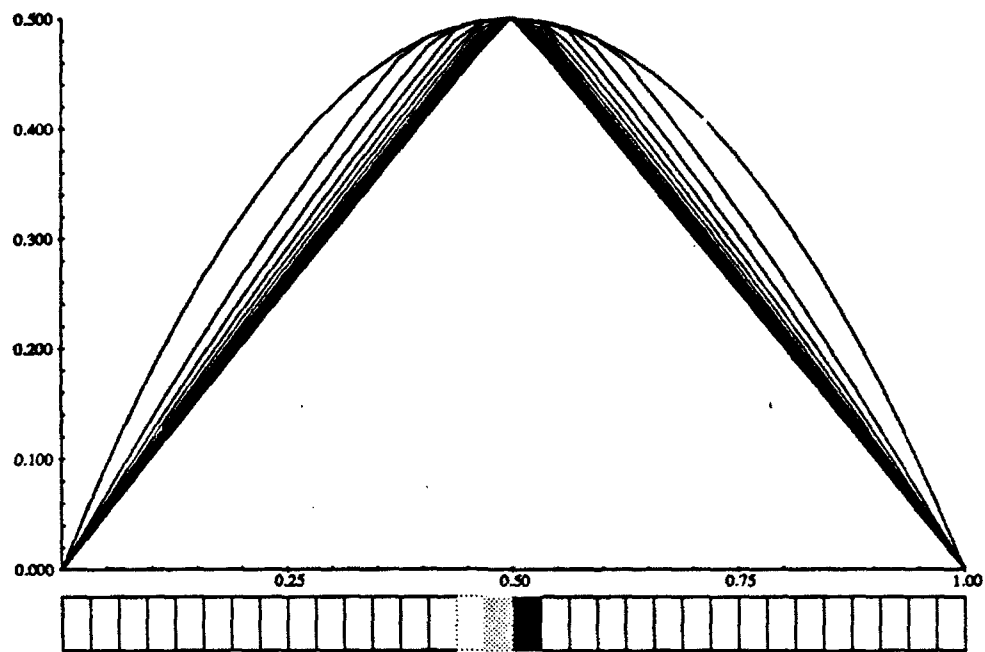
$$\text{supp } v^{h,k} \subset \{\lambda: \varphi(\lambda) = \varphi^{**}(\lambda)\}, \quad (5.5)$$

where φ^{**} denotes the convexified φ , which is its relaxation in this situation. Moreover, it is possible to show that

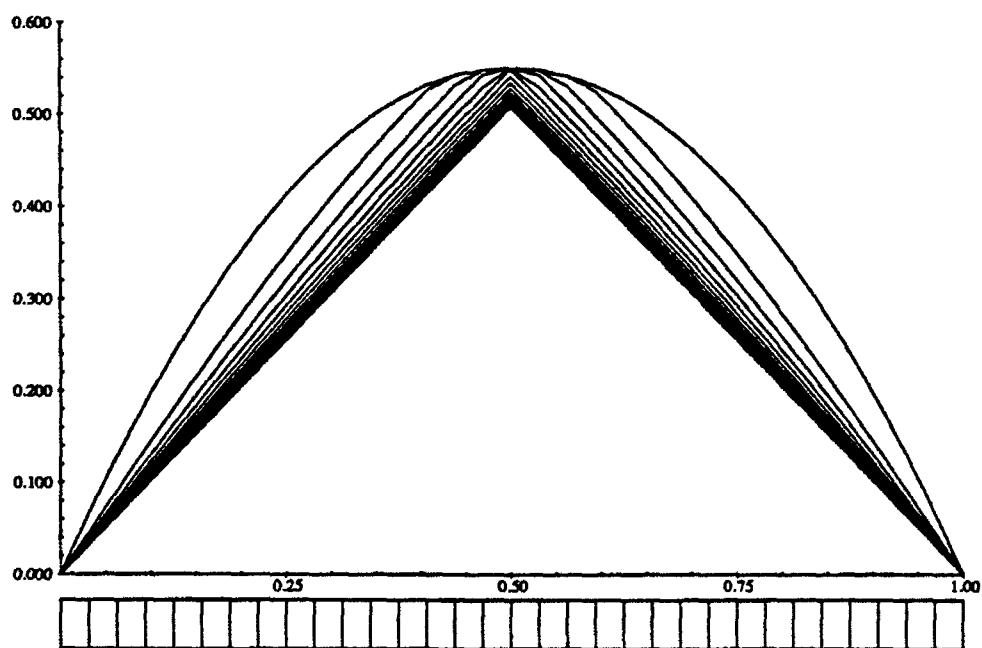
$$-\text{div } \bar{q}^h + \frac{1}{h}(u^{h,k} - u^{h,k-1}) = 0, \quad \text{where} \quad (5.6)$$

$$\bar{q}^h(x) = \int_{\mathbb{R}^N} q^h(\lambda) dv_x^{h,k}(\lambda).$$

We next assemble the $v^{h,k}$ and $u^{h,k}$, defining u^h to be the linear interpolant of the $(u^{h,k})$ and v^h the piecewise constant in time measure equal to $v^{h,k}$ in $(k-1)h < t \leq kh$. We obtain in this fashion a family of approximants which are "maximally dissipative" because of (5.5) and converge to a Young measure solution of (5.1).



$$u(x,0) = 2x(1-x)$$



$$u(x,0) = 2.2x(1-x)$$

Figure 6 Computation of the solution of (5.6) by a Young measure algorithm.

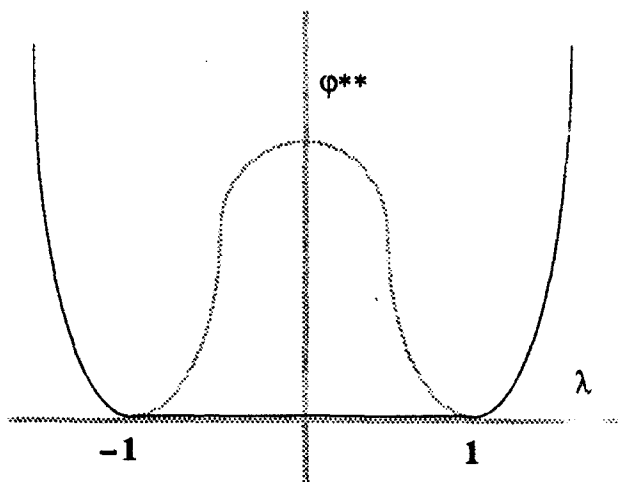


Figure 5. The relaxation ϕ^{**} of ϕ plays an important role in the solution of the problem. Where ϕ is different from ϕ^{**} is shown in dotted lines.

It turns out that the solution $u^{h,k}$ of the basic variational principle is the unique solution of the relaxed problem for this principle which is strictly convex in v , but not in ∇v . Some convex analysis then tells us that u^h is unique and so is its limit u as $h \rightarrow 0$. The Young measure need not be unique.

This is an opportunity to study the pure generation of oscillations in a new context. The underlying solution is known. What statistics are possible for the sequences

which generate this solution and can we compute them? How do they evolve in time? Using an algorithm specifically designed to compute Young measures by Nicolaides and Walkington, Walkington has computed several examples of this behavior, shown in Figure 6. Although in these pictures, the solution decreases monotonically to its limit

$$u(x, \infty) = \begin{cases} x & 0 < x \leq 0.5 \\ 1 - x & 0.5 < x < 1 \end{cases},$$

this is not the case for general initial values $u(x, 0)$, although the energy integral is a decreasing function of t . We refer also to their article in these proceedings.

The author wishes to thank his colleagues and coworkers, especially R. James, R. Nicolaides, P. Pedregal, and Noel Walkington, for their assistance and advice in preparing this account.

References

- [1] Ball, J. M. and James, R. 1987 Fine phase mixtures as minimizers of energy, *Arch. Rat. Mech. Anal.*, 100, 15-52
- [2] Ball, J. M. and James, R. 1991 Proposed experimental tests of a theory of fine microstructure and the two well problem, *Phil. Trans. Roy. Soc. Lond.* (to appear)
- [3] Barrett, C. and Massalski, T. B. 1980 *Structure of Metals*, 3rd rev. ed., Pergamon
- [4] Battacharya, K. Self accommodation in martensite, *Arch. Rat. Mech. Anal.* (to appear)
- [5] Battacharya, K. Wedge-like microstructure in martensite, (to appear)

- [6] Brown, W.F. 1962 *Magnetostatic Principles in Ferromagnetism*, Vol. 1 of Selected Topics in Solid State Physics (ed. E.P. Wohlfarth), North-Holland.
- [7] Brown, W.F. 1963 *Micromagnetics*, John Wiley and Sons, New York.
- [8] Brown, W.F. 1966 *Magnetoelastic Interactions*, Vol. 9 of Springer Tracts in Natural Philosophy (ed. C. Truesdell), Springer-Verlag.
- [9] Chipot, M. and Kinderlehrer, D. 1988 Equilibrium configurations of crystals, *Arch. Rat. Mech. Anal.* 103, 237-277
- [10] Chipot, M. Numerical analysis of oscillations in nonconvex problems, (to appear)
- [11] Clark, A. E. 1980 Magnetostrictive rare earth - Fe_2 compounds, *Ferromagnetic Materials, Vol 1* (Wohlfarth, E. P. ed) North Holland, 532 - 589
- [12] Collins, C. and Luskin, M. 1989 The computation of the austenitic-martensitic phase transition, *Lecture Notes in Physics* 344 (ed. M. Rascle, D. Serre and M. Slemrod), Springer-Verlag, 34-50.
- [13] Collins, C. and Luskin, M. Numerical modeling of the microstructure of crystals with symmetry-related variants, *Proc. ARO US-Japan Workshop on Smart/Intelligent Materials and Systems*, Technomic
- [14] Collins, C. and Luskin, M. Optimal order error estimates for the finite element approximation of the solution of a nonconvex variational problem, to appear
- [15] Collins, C., Kinderlehrer, D., and Luskin, M. 1991 Numerical approximation of the solution of a variational problem with a double well potential, *SIAM J. Numer. Anal.*, 28, 321-333
- [16] Ericksen, J. L. 1979 On the symmetry of deformable crystals, *Arch. Rat. Mech. Anal.* 72, 1-13
- [17] Ericksen, J. L. 1980 Some phase transitions in crystals, *Arch. Rat. Mech. Anal.* 73, 99-124
- [18] Ericksen, J. L. 1981 Changes in symmetry in elastic crystals, *IUTAM Symp. Finite Elasticity* (Carlson, D.E. and Shield R.T., eds.) M. Nijhoff, 167-177
- [19] Ericksen, J. L. 1982 Crystal lattices and sublattices, *Rend. Sem. Mat. Padova*, 68, 1-9
- [20] Ericksen, J. L. 1984 The Cauchy and Born hypotheses for crystals, *Phase Transformations and Material Instabilities in Solids*, (Gurtin, M., ed) Academic Press, 61-78
- [21] Ericksen, J. L. 1986 Constitutive theory for some constrained elastic crystals, *Int. J. Solids Structures*, 22, 951 - 964
- [22] Ericksen, J. L. 1987 Twinning of crystals I, *Metastability and Incompletely Posed Problems*, IMA Vol. Math. Appl. 3, (Antman, S., Ericksen, J.L., Kinderlehrer, D., Müller, I., eds) Springer, 77-96
- [23] Ericksen, J. L. 1988 Some constrained elastic crystals, *Material Instabilities in Continuum Mechanics*, (Ball, J. ed.) Oxford, 119 - 136
- [24] Ericksen, J. L. 1989 Weak martensitic transformations in Bravais lattices, *Arch. Rat. Mech. Anal.* 107, 23 - 36
- [25] Firooze, N. and Kohn, R. 1991 Geometric parameters and the relaxation of multiwell energies, *IMA preprint Series* 765
- [26] Fonseca, I. 1988 The lower quasiconvex envelope of the stored energy function for an elastic crystal, *J. Math. pures et appl.* 67, 175-195
- [27] James, R. D. 1988 Microstructure and weak convergence, *Proc. Symp. Material Instabilities in Continuum Mechanics*, Heriot-Watt, (Ball, J. M., ed.), Oxford, 175-196
- [28] James, R. D. 1989 Relation between microscopic and macroscopic properties of crystals undergoing phase transformation, in *Proc. 7th Army Conf. on Applied Mathematics and Computing* (ed. F. Dressel).
- [29] James, R. D. and Kinderlehrer, D. 1989 Theory of diffusionless phase transitions, *PDE's and continuum models of phase transitions*, *Lecture Notes in Physics*, 344, (Rascle, M., Serre, D., and Slemrod, M., eds.) Springer, 51-84
- [30] James, R. D. and Kinderlehrer, D. 1990 An example of frustration in a ferromagnetic material, *Nematics: Mathematical and Physical Aspects*, (Coron, J.-M., Ghidaglia, J.-M., and Hélein, F., eds), Kluwer NATO ASI series, 201-222
- [31] James, R. D. and Kinderlehrer, D. 1990 Frustration in ferromagnetic materials, *Cont. Mech. Therm.* 2, 215-239
- [32] James, R. D. and Kinderlehrer, D. A theory of magnetostriction with application to TbDyFe_2 (to appear)
- [33] James, R.D. and Müller, S. to appear
- [34] Khachaturyan, A. G. 1983 *Theory of structural phase transformations*, Wiley
- [35] Kinderlehrer, D. 1988 Remarks about the equilibrium configurations of crystals, *Proc. Symp. Material instabilities in continuum mechanics*, Heriot-Watt (Ball, J. M. ed.) Oxford, 217-242

- [36] Kinderlehrer, D. and Pedregal, P. 1991 Characterizations of Young measures generated by gradients, *Arch. Rat. Mech. Anal.*, 115, 329-365
- [37] Kinderlehrer, D. and Pedregal, P. 1991 Caractérisation des mesures de Young associées à un gradient, *C.R.A.S. Paris*, 313, 765-770
- [38] Kinderlehrer, D. and Pedregal, P. 1992 Gradient Young measures generated by sequences in Sobolev spaces (to appear)
- [39] Kinderlehrer, D. and Pedregal, P. 1992 Remarks about the analysis of gradient Young measures, *Pitman Res Notes Math* 269 (Miranda, M., ed), 125-150
- [40] Kinderlehrer, D. and Pedregal, P. 1992 Weak convergence of integrands and the Young measure representation, *SIAM J. Math Anal.*, 23, 1 - 19
- [41] Kinderlehrer, D. and Pedregal, P. Remarks about Young measures supported on two wells, (to appear)
- [42] Kohn, R. V. 1991 The relaxation of a double-well energy, *Cont. Mech. Therm.*, 3, 193-236
- [43] Kohn, R. V. 1989 The relationship between linear and nonlinear variational models of coherent phase transitions, in *Proc. 7th Army Conf. on Applied Mathematics and Computing* (ed. F. Dressel).
- [44] Lord, D. 1990 Magnetic domain observations in TbDyFe₂, IMA lecture
- [45] Luskin, M. and Ma, L. 1990 Analysis of the finite element approximation of microstructure in micromagnetics, UMSI report 90/164
- [46] Ma, L. to appear
- [47] Matos, J. The absence of fine microstructure in α - β quartz, (to appear)
- [48] Nicolaides, R. and Walkington, N. these proceedings
- [49] Roitburd, A. L. 1978 Martensitic transformation as a typical phase transformation, *Solid State Phys* 33, Academic Press, 317-390
- [50] Tartar, L. 1979 Compensated compactness and applications to partial differential equations, *Nonlinear analysis and mechanics: Heriot Watt Symposium, Vol I V* (Knops, R., ed.) Pitman Res. Notes in Math. 39, 136-212
- [51] Tartar, L. 1983 The compensated compactness method applied to systems of conservation laws, *Systems of nonlinear partial differential equations* (Ball, J. M., ed) Riedel
- [52] Tartar, L. 1990 H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations, *Proc. R. Soc. Edin.*, 115, 193-230
- [53] Teter, J.P., Mahoney, K., Al-Jiboory, M., Lord, D., and McMasters, O. D., 1991 Domain observation and magnetostriction in TbDyFe twinned single crystals, *J. Appl. Phys.*, 69, 5768-5770
- [54] Toupin, R. 1956 The elastic dielectric, *J. Rat. Mech. Anal.*, 5, 849 - 915
- [55] Young, L. C. 1969 *Lectures on calculus of variations and optimal control theory*, W.B. Saunders

Classical Finite Element Method for Transient
Three Dimensional Heat Conduction

Rao Yalamanchili and S. Yalamanchili*
Propulsion Branch, Energetics and Warheads Division
Armament Engineering Directorate
U.S. Army Armament Research, Development and Engineering Center
Picatinny Arsenal, NJ 07806-5000

ABSTRACT. Various finite difference schemes were presented in the Ninth Army Conference on Applied Mathematics and Computing held at University of Minnesota Super Computing Center. It is of interest to know where classical finite element method stands against its counterpart. Towards this goal, the finite difference equivalent is derived by the use of Gurtin's variational principle, introduced during mid 60s, for linear initial value problems and finite element synthesis. The body is divided into various finite elements. Linear temperature distribution within each element and linear temperature variation within each time step are allowed as in classical finite element method. The resulting equation is found to be same as one of the finite difference equation derived earlier, i.e., equal weights at all nodes of a 27-node element except at the center. This is not the best one as far as accuracy, stability and nonoscillation characteristics are concerned. A question is now posed to the readers about consequences of liberalizing the two basic assumptions of classical finite element method.

INTRODUCTION. The motivation for this task comes from the needs of another project, aerodynamic heating of hypervelocity projectiles. The transient three dimensional heat conduction model will provide a means to determine the temperature distribution as a function of location (three dimensions) and time for any given initial and boundary conditions. The boundary conditions are usually obtained from computational fluid dynamics models. There are occasions where there is a strong coupling between convective flow and conduction. No matter what method is utilized, all require large computer storage and a great amount of computer time. The solution process is not only subjected to these restrictions but also bound to blow-up, in the middle, if proper selection of numerical techniques and accuracy, stability and nonoscillation characteristics are not taken into account. If a combined convection and conduction problem is attempted in one-step, the failure in one area can lead to a losing proposition in both areas. Therefore, a search is initiated to find an accurate, robust and efficient numerical scheme for the solution of transient three dimensional heat conduction problems.

*College of Engineering, Rutgers University, New Brunswick, NJ

VARIATIONAL PRINCIPLE. Wilson and Nickel [1], following Gurtin's [2] discussion of variation principles for linear initial value problems, confirmed that the function $T(x,y,z,t)$ which leads to an extreme of functional

$$\Omega_e(T) = 1/2 \int_V \{ \rho C_p T * T + \nabla T * K * \nabla T - 2 \rho C_p T_0 * T \} dV \\ - \int_S \hat{Q}_i n_i * T dS$$

is the solution of the following transient heat-conduction equation:

$$(K * T, i), i - \rho C_p * \frac{\partial T}{\partial t} + \rho * p = 0$$

$$\text{with the boundary condition } K * T, i - \hat{Q}_i = 0$$

Where $T(x,y,z,t)$ = temperature at the spatial point (x,y,z) and at time t

T_0 = Initial temperatures

T = Gradient of T with respect to spatial coordinates

K = Thermal conductivity

ρ = Material density

C_p = Heat capability of the material per unit mass

$$\hat{Q}_i(x,y,z,t) = \int_0^t Q_i(x,y,z,\tau) d\tau$$

U = Volume

$*$ = Convolution symbol defined as:

$$T * T = \int_0^t T(x,y,z,t-\tau) T(x,y,z,\tau) d\tau$$

$$\nabla T * \nabla T = \frac{\partial T}{\partial x} * \frac{\partial T}{\partial x} + \frac{\partial T}{\partial y} * \frac{\partial T}{\partial y} + \frac{\partial T}{\partial z} * \frac{\partial T}{\partial z}$$

FINITE ELEMENT SYNTHESIS. Divide the three dimensional solid body into I axial elements (nodes 0 to I in x -direction), J transverse elements (nodes 0 to J in y -direction), and K normal elements (nodes 0 to K in z -direction) such that step sizes are same in all three directions. This restriction is introduced to simplify algebraic manipulations involved in the analysis. Also, a unit step size is assumed to simplify derivations and generalized later.

Consider the nodal point (i,j,k) , in the range $0 < i < I$, $0 < j < J$, $0 < k < K$ as shown in Figure 1. The temperature of the nodal point will vary as a function of time, t . The temperature distribution in a subregion is a function of spatial coordinates (x,y,z) and surrounding nodal point temperatures. For simplicity, linearity and the same functional distribution are assumed for all elements. The functions $f_1, f_2, f_3, f_4, f_5, f_6, f_7$, and f_8 are functions of nodal point temperatures. These are determined by substitution of the coordinates of nodal points into the equation and by solving the resulting simultaneous equations. The results for region II are given on the next page.

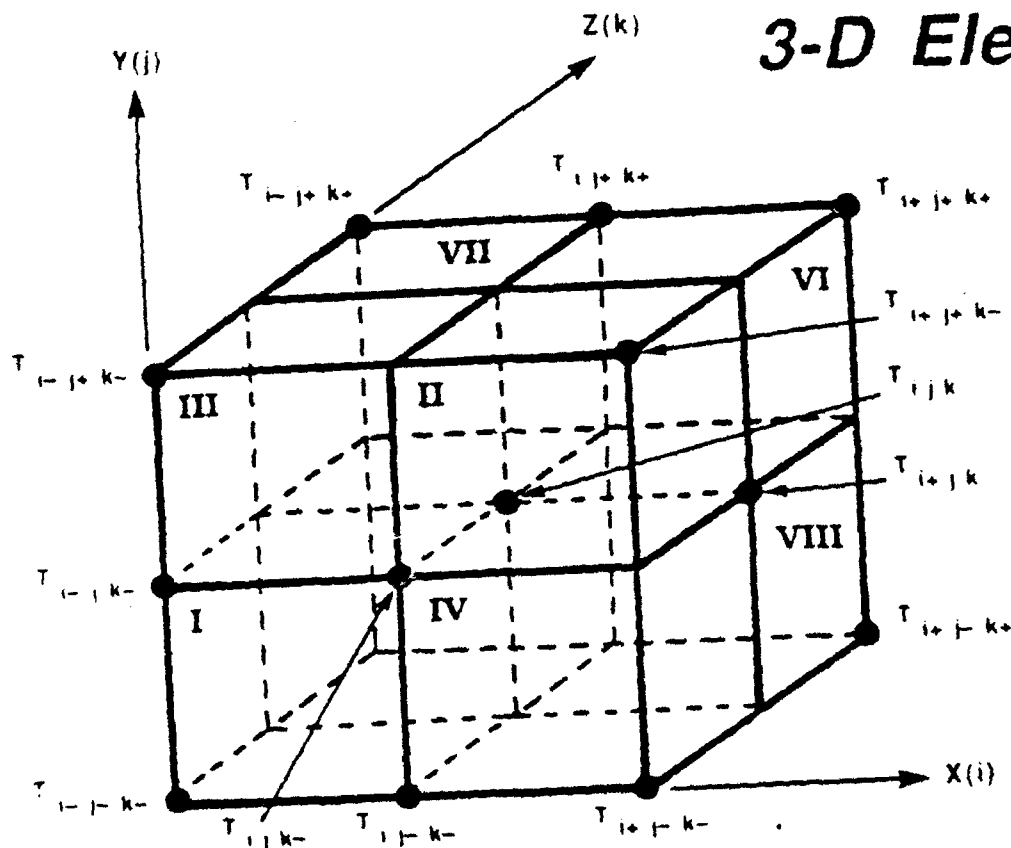


Figure 1

Region II:

$$T(x, y, z) = f_1 + f_2 x + f_3 y + f_4 z + f_5 xy + f_6 xz + f_7 yz + f_8 xyz$$

$$f_1 = T_{ijk} = T_{ooo}$$

$$f_2 = T_{+oo} - T_{ooo}$$

$$f_3 = T_{o+o} - T_{ooo}$$

$$f_4 = T_{ooo} - T_{oo-}$$

$$f_5 = T_{++o} - T_{+oo} - T_{o+o} + T_{ooo}$$

$$f_6 = T_{+oo} - T_{+o-} - T_{ooo} + T_{oo-}$$

$$f_7 = T_{o+o} - T_{o+-} - T_{ooo} + T_{oo-}$$

$$f_8 = T_{++o} - T_{++-} - T_{+oo} + T_{+o-} - T_{o+o} + T_{o+-}$$

Where the subscripts + and - denote nodes one step ahead and one-step behind respectively. Once the temperature distribution is obtained, anyone can derive the partial first order derivatives from the above equation. Similar set of equations can be derived for the remaining seven regions of the three dimensional finite element, discussed above. It is time now to substitute all equations derived into all three terms of functional, governing equation, and integrate over the volume occupied by region II and take the first variation with respect to T_{ijk} at the new time in order to obtain the extremum of the functional. Obtain similar results for the other regions and combine them.

The procedures for the first and the third term of the governing equation are same. However, double convolution symbol is involved in the second term. Here the evaluation of the second term in the governing equation involves integration not only over the volume (and the use of first variation with respect to the nodal temperature, T_{ijk}^+ at the new time) but also over the time-step due to the additional convolution symbol. Towards this goal, a linear nodal point temperature variation is assumed within each time-step. Summing up the results of all three terms produces the following finite element equivalent:

The superscripts + and o denote new time and old time, respectively, for any time-step. θ is the dimensionless parameter, Fourier number ($\frac{1}{2} \Delta t / \Delta x^2$). The left side of this equation represents the weighted average of time-derivatives at all nodes of a 27-node finite element system. Of course, the weights are different depending upon the location of the node. The right side of this equation is equivalent to an arithmetic average of 3-dimensional Laplacian (in a particular finite difference form) at two different times, old and new. This particular finite difference form for representing a laplacian is not the best one as far as accuracy and nonoscillation characteristics are concerned [3]. The details of derivation that lead to finite element equivalent are omitted here due to lack of space and time. However, it is planned to publish in a journal, similar to the two-dimensional case [4].

One can also rewrite the above equation, finite element equivalent, in a more familiar form to the finite difference community as shown below. All unknown nodal point temperatures are arranged on the left side and all known quantities are shuffled to the right side of this equation just like in a traditional way. The coefficients, A through H, are different. These are not given here due to lack of space. However, one can write them by comparison of this equation with the above finite element equivalent.

$$\begin{aligned}
 & AT_{...}^{+} + BT_{...o}^{+} + AT_{...+}^{+} + BT_{...o}^{+} + CT_{...oo}^{+} + BT_{...o}^{+} + \\
 & AT_{...+}^{+} + BT_{...+o}^{+} + AT_{...++}^{+} + BT_{...o++}^{+} + CT_{...o+o}^{+} + BT_{...o++}^{+} + \\
 & CT_{...oo+}^{+} - DT_{...ooo}^{+} + CT_{...oo+}^{+} + BT_{...o++}^{+} + CT_{...o+o}^{+} + BT_{...o++}^{+} + \\
 & AT_{...+}^{+} + BT_{...+o}^{+} + AT_{...++}^{+} + BT_{...o+}^{+} + CT_{...+oo}^{+} + BT_{...+o+}^{+} + \\
 & AT_{...++}^{+} + BT_{...++o}^{+} + AT_{...+++}^{+} = ET_{...}^{o} + FT_{...o}^{o} + ET_{...+}^{o} + \\
 & FT_{...o}^{o} + GT_{...oo}^{o} + FT_{...o+}^{o} + ET_{...++}^{o} + FT_{...+o}^{o} + ET_{...++}^{o} + \\
 & FT_{...o+}^{o} + GT_{...o+o}^{o} + FT_{...o++}^{o} + ET_{...+++}^{o} + FT_{...+o+}^{o} + ET_{...+++}^{o} + \\
 & FT_{...o+}^{o} + GT_{...+oo}^{o} + FT_{...+o+}^{o} + ET_{...+++}^{o} + FT_{...++o}^{o} + ET_{...+++}^{o}
 \end{aligned}$$

OTHER FINITE ELEMENT SCHEMES. Classical finite element scheme is discussed above. Linear temperature distribution, in space, is allowed in addition to linear temperature variation within each time-step. Similarly, one can also treat quadratic, cubic, etc., variation in space. However, analytical evaluation, similar to the above, is almost impossible because of tedious and extensive algebraic work and associated errors involved. Sometimes, this is referred to in the literature as h-version finite element method.

There are three well established general approaches in the finite element method: h-version, p-version and h/p-version. In the h-version of finite element method, the element approximation is kept fixed and the mesh is refined thereby adding more degrees of freedom to improve the accuracy of the solution. In the p-version of finite element method, the order of approximation for each element of a fixed grid is increased thereby adding more degrees of freedom to an existing model to improve the accuracy of the solution. In the h/p-version of the finite element, the mesh refinement and the order of approximation for the element is accomplished simultaneously. The p-version of finite element methods yields a much higher convergence rate and requires very simple models which need to be constructed only once. For singular problems such as near freezing or melting interfaces, the rate of convergence of the p-version is even higher than for smooth problems. Computational benefits of hierarchical properties result in additional saving and efficiency. The inability of h-version to correctly simulate singular problem behavior is well known. In such applications p-version (or h/p-version) is the only feasible approach. As a result this newer technique has found considerable application in modeling composite structures, large strain and deformation.

CONCLUSIONS. Hypervelocity projectile is in the DODs critical technology list. The task reported here directly supports this project. Classical finite element analysis is performed on transient three dimensional heat conduction problems by the use of variational principle and finite element synthesis. The finite element equivalent, derived here, appears first time in the literature. The terms on the left and right sides are identified. It is found that the classical finite element analysis may not be good based on physical interpretation of terms and comparison with other finite difference schemes especially if accuracy, stability and nonoscillation characteristics are taken into account. Other finite element variations which are superior at least for special problems such as melting or freezing are also discussed if one is interested in pursuing only finite element methodology.

REFERENCES.

1. Wilson, E.L., and R.E. Nickell, "Application of the Finite Element Method to Heat Conduction Analysis," Nuclear Engineering and Design, Vol. 4, pp. 276-286 (1966).
2. Gurtin, M.E., "Variational Principles for Linear Initial Value Problems," Quarterly Applied Mathematics, Vol. 2, pp. 252-256 (1964).

3. Yalamanchili, R., and S. Yalamanchili, "Various Finite Difference Schemes for Transient Three Dimensional Heat Conduction," Transactions of the Ninth Army Conference on Applied Mathematics and Computing, ARO Report 92-1 (March 1992).
4. Yalamanchili, R., and S. Chu, "Stability and Oscillation Characteristics of Finite Element, Finite Difference and Weighted Residual Methods for Transient Two-Dimensional Heat Conduction in Solids'" Journal of Heat Transfer, Trans. of ASME, Vol. 95, Series C, #2 (1973).

NON-COLLOCATED MOTION CONTROL OF A FLEXIBLE BEAM BASED ON A DELAYED ADAPTIVE INVERSE METHOD¹

David S. Wang, Guo-Ben Yang, Max Donath
Army High Performance Computing Research Center,
and the Department of Mechanical Engineering
University of Minnesota
111 Church Street S.E.
Minneapolis, Minnesota 55455

Mike Mattice and Norman Coleman, Jr.
U.S. Armament Research,
Development, and Engineering Center
(ARDEC)
Picatinny Arsenal, New Jersey 07801

Abstract

Motion control of mechanical systems with inherent structural and joint flexibility represents a class of nonlinear systems found in many important applications. Applications include the motion control of machines used widely in weapons, medicine, construction, mining, and manufacturing. Control of the tip motion (e.g., the end effector in the case of a robot manipulator or the active end of a pointing system) in an inertial coordinate frame is the objective in most such applications. This is the case because tasks are usually specified by the desired motion trajectory of the tip described inertially. As such, one needs to compute and control the applied actuator torques so that the motion of the tip will match the prescribed motion. However, a non-minimum phase system arises when one attempts to control the motion of a mechanically compliant system by using a measurement of the tip to control a non-collocated actuator.

The viability of a delayed adaptive inverse method for motion control of the tip of a flexible beam is demonstrated in this paper. This method is based on an adaptive linear FIR filter which provides a stable and close approximation to the inverse plant dynamics. Such FIR filters can be used to control non-minimum phase systems, certain nonlinear systems, or plants of unknown dynamics and can be implemented using real-time interrupt driven high performance computational devices such as digital signal processing (DSP) based hardware. A least mean square (LMS) error minimization is used to update the weight vector which forms a delayed, adaptive inverse dynamic model of the system. By using an FIR filter to represent the inverse of plant dynamics, the instability associated with pole zero cancellation on the right half Laplace plane is avoided when controlling non-minimum phase systems. Properly selected values of the initial weights and of the gain constant, μ , guarantee a bounded solution for the computed input torque. The effects of varying parameters of the delayed adaptive inverse controller (such as the values of the initial weights, the gain constants, and the length of delay) on the system performance are presented. Since the method is computationally efficient, it can be used in high bandwidth applications for the control of complex plants given only the desired and the actual measured outputs for the plant. A simple feedback controller is used in conjunction with the delayed adaptive inverse controller to eliminate the residual errors that would occur if using only feed forward control.

The delayed adaptive inverse controller was designed based on a model of the flexible beam test bed located at the Armament Research, Development and Engineering Center (ARDEC). The performance of this adaptive controller is compared with that of an H_{∞} based loopshaping controller. The results indicate that higher performance is achieved when using the adaptive inverse controller if sensor noise is minimal. If not, the robust H_{∞} controller would be recommended.

¹ Work was partially supported by the Army High Performance Computing Research Center and the Productivity Center at the University of Minnesota.

1. Introduction

There are many tasks in which the precise motion control of a point on a moving structure is of interest rather than the actuator's motion itself. Robot manipulators are but one example of such a system. Each robot's actuated link can be considered to be a motor driven beam in which the beam's tip motion must be controlled. For manipulators, tasks are specified by the desired motion trajectory of the tip. This means that one needs to compute and control the applied torque at each joint so that the measured motion of the tip will match the prescribed motion. Two problems arise in this case. One, sensors for measuring tip motion have to date, been severely limited in bandwidth; and two, given the non-minimum phase nature of such non-collocated systems, the use of traditional inverse dynamics based control will lead to instabilities². The problem in our case is further exacerbated by the nonlinear nature of the plant to be controlled. Nonlinearities here include the inherent geometric/continuous nonlinearities and the discontinuity effects such as backlash, stiction and friction. An exact Fourier series-based inverse method has been developed (Yang, 1991), in which good agreement was achieved between the desired and computed motion. However, the use of such exact inverse dynamics based formulations for the control of non-minimum phase systems requires prior knowledge regarding the desired motion/path boundary conditions and can lead to instabilities due to the generation of unstable poles in the controller.

In order to address the first problem above, we have developed a sensor that can be used for tracking the six degrees of freedom motion of an end effector. In Sorensen et al. (1989), we described our first prototype which is capable of high bandwidth (480 Hz) measurement of the XYZ coordinates of moving points on rigid or deformable bodies. The system, based on a laser scanning approach, is designed in a pipeline configuration such that each sensed point's coordinates are immediately available in registers and memory mapped into the data acquisition CPU. Three planes of light rotate through the measurement field at constant angular velocity. By measuring the elapsed time for rotation from fixed locations at the boundaries of the field to each of the moving point targets, one can derive the swept angles to each of the targets and consequently their XYZ coordinates. In addition to being able to track each point target's motion, each target has its own path into the CPU thus facilitating the accurate computation of (a) the six degrees of freedom of any number of bodies each carrying at least three targets and (b) the relative motion of the instantaneous axis of rotation between bodies. Significant improvements have been implemented as a follow-up to the experiments described in Sorensen et al. (1989). As a result, such sensors can now be used to characterize the relative motion of joints (e.g. transmissions) and structural elements even when compliance is a major factor. It is now therefore possible to continuously and simultaneously monitor the motion of one or more mobile base platforms, the exact behavior of each of the robot joints, their motors and transmission effects, the structural deformations of the links, and the full six degrees of freedom of the end effector (or load) using one sensing system. This advance in technology makes it possible to consider multi-degree of freedom end point control.

In this paper, we will focus on a method to handle the second problem mentioned earlier, that of controlling a nonlinear non-minimum phase system. We will present a method for controlling the tip motion of a flexible beam. In contrast to most inverse dynamics methods proposed by other researchers, the delayed adaptive inverse method that we propose attempts to address the problems associated with the non-minimum phase behavior of the flexible system and provides a stable and close approximation to the inverse dynamics. Furthermore, the approach is conducive to the use of high speed Digital Signal Processing (DSP) hardware. Most of the early research work on flexible robot manipulators used collocated feedback methods. As such, no direct measurement was made of the effect of compliance or other unexpected dynamics located between the actuator and the point

² A non-minimum phase system typically arises when one attempts to control the motion of a flexible structure using a measurement of the tip to control a non-collocated actuator.

of interest (e.g. manipulator tip) under control. Several papers falling into this category are discussed below.

Book et al. (1975) implemented a flexible feedback control scheme (FFC) based on a pole shifting algorithm. Siciliano and Book (1988) applied the singular perturbation and integral manifold approaches to control a single link flexible arm. To overcome the sensitivity of these methods to plant parameter variations, adaptive methods were investigated. Nelson and Mitra (1986) used on-line load mass estimation and load-adaptive optimal control for varying load conditions. Other adaptive control methods were also proposed by Meldrum and Balas (1985). However all these adaptive control designs involve a significant computational component thereby limiting their application to real time control. Again these papers are limited to cases of collocated feedback.

To obtain better performance for very flexible manipulators, Cannon and his research group (Cannon and Schmitz, 1984; Schmitz, 1989) experimentally investigated end-point control techniques for flexible link systems in planar motion. This was the first attempt at non-collocated control for a flexible beam model of a manipulator using traditional control design methodologies. Lee and Castelazo (1987) developed a non-collocated sensor-actuator scheme with non-linear feedback to control a flexible manipulator. However, there were a number of limitations associated with this latter group of non-collocated techniques, such as the inability to handle joint compliance and the uncertainty associated with a model of the system.

Research on the control of systems with joint flexibility was first initiated in the 1980s. Under the assumption of weak elasticity at the joint, Ficola et al. (1983) used the singular perturbation approach to design a feedback controller. Khorasani and Spong (1985) extended this work by using the invariant and integral manifold approach. A pseudo-linearization technique was proposed by Nicosia et al. (1986) while adaptive control approaches were investigated by Tomei et al. (1986). All these studies ignored structural compliance and cannot readily be modified to incorporate such compliance.

Furthermore, the control methods in most of the literature mentioned above was based on an analysis of the forward dynamics, in which the flexible link (or joint) displacements, velocities and accelerations are determined by the specified joint torque. An inverse dynamics type of control law based on the concept of "Computed Torque" was first applied to a cylindrical coordinate arm with drive train compliance by Forrest-Barlach and Babcock (1986). Other such feedforward methods followed. Asada and Ma (1989) presented a recursive inverse dynamics analysis based on a virtual rigid link coordinate system. Tsujio (1988) suggested another approach, in which the driving force was calculated based on the rigid manipulator, and then applied to a flexible manipulator iteratively until it converged. None of this latter group of inverse based methods was adaptive and as such are sensitive to parameter uncertainty. Furthermore, none took advantage of tip feedback and suffered from the limitations inherent to systems in which direct measurement of the parameter under control is not available.

An approach that explicitly deals with the non-minimum component of a plant was described by Tomizuka (1987) who proposed a zero phase error tracking controller (ZPETC) to solve a tracking control problem. This approach used a feedforward controller to cancel all closed-loop poles and cancelable closed-loop zeros (zeros on the left-half of the s -plane). For uncancelable zeros (zeros on the right-half of the s -plane), the feedforward controller cancels the phase shift induced by them. Since ZPETC is based on pole/zero cancellation and phase cancellation, the tracking performance using ZPETC is sensitive to modeling errors and plant parameter variations. To solve the problem with unknown plant parameters and parameter variations, Tsao and Tomizuka (1987) developed an adaptive ZPETC. The system plant is separated into a known part (with cancelable poles and zeros) and an unknown part. A normalized least square parameter adaptation algorithm (PAA) was used to adjust the unknown parameters used in the feedforward

controller. However, in both cases, the results were dependent on having a good model of the system.

Another approach for designing controllers of non-minimum phase, linear systems is H_∞ based loopshaping (Doyle et al., 1992). This approach presents a graphical technique for designing a controller to achieve robust performance, i.e., performance in the presence of uncertainty. The idea of loopshaping is to shape the complementary sensitivity transfer function, the transfer function from the reference signal to the output of the plant, in order to obtain the desired system performance. The ideal loopshape of the complementary sensitivity transfer function is represented by a second order system with desired response characteristics such as the amount of damping and natural frequency. This transfer function is then incorporated, as a weighting function, into the system model for a controller design that minimizes the induced two norm between the reference signal and the output of the weighting function.

In this paper we examine a method based on a delayed adaptive inverse filter which eliminates the non-minimum problem associated with inverse methods by using an FIR filter for feedforward control. The method is based on the on-line characterization of the inverse dynamics based on an adaptation algorithm in which a delay must be injected to ensure causality. We will demonstrate the viability of the delayed adaptive inverse method for motion control of flexible beams. An adaptive linear function whose weighting is adapted by the LMS algorithm was used to replace the unstable exact inverse model. By using a FIR filter to represent the inverse dynamics, we avoid the instabilities associated with pole zero cancellation (only zeros are possible in FIR filters) when controlling non-minimum phase systems, and we can now take advantage of high speed computational architecture based on the DSP chip. Furthermore, we may be able to control complex plants given only the desired and actual output values for the plant, provided that the control inputs are accessible. Since the method is computationally efficient, it can be incorporated into closed-loop controllers for position or trajectory control so that the residual errors can be eliminated (see discussion later). It can also be applied to multi-degree-of-freedom systems.

This paper significantly expands on an earlier preliminary presentation describing the method (Yang and Donath, 1990). A general description of the delayed adaptive inverse model is provided in section 2. The weight vector used in this model is updated using the LMS algorithm, which is discussed in section 3. The contents of sections 2 and 3 were presented in Yang and Donath (1990), but they are included here for completeness. A control strategy based on the adaptive inverse model is described in section 4. The approach, as applied to a model of the flexible beam located at the Armament Research, Development and Engineering Center, is described in section 5. A comparison of this adaptive inverse controller with an H_∞ based loopshaping controller is then discussed in section 5.

2. A Delayed Adaptive Inverse Model

Adaptive filters are widely used in digital signal processing, such as in communications, radar, sonar, and seismology, where a priori data information is unknown (Haykin, 1986). In a stationary process, the algorithm starts with a set of initial conditions and converges to its optimum solution after successive iterations. In the nonstationary case, the algorithm tracks time variations in the statistics of the input data for a slow changing process. The application of adaptive filters to control design can be categorized into two groups: (i) the adaptive model control (AMC) method; and (ii) the adaptive inverse control (AIC) method. It is difficult to control a non-minimum phase system by using the AMC method, since the control signal will have a transform with poles outside the unit circle and will thus be unstable. However, a delayed AIC method can form stable approximate inverse models without knowing a priori whether or not the plant is minimum phase.

In this paper, we will apply the delayed AIC method to the tip control of flexible beams, a non-minimum phase system. An adaptive linear combiner, or nonrecursive adaptive filter, is used to

compute the applied joint torque, which, in turn, generates a desired motion. One advantage, among many, of the adaptive linear combiner is that it is relatively straightforward to analyze and implement. A schematic diagram of a delayed adaptive inverse model used in the AIC method is shown in Figure 1, where the terms, u_k , d_k , n_k , Y_k , V_k , and ϵ_k are respectively the command input, the command after a delay of Δ sample periods, the noise input, the plant output, the output of the inverse plant model, and the error between d_k and V_k . $P(z)$ and $H(z)$ are z transfer functions of the plant and of the adaptive inverse model.

This inverse modeling approach, known as adaptive equalization in the communications field, was first used in the 1960s to counter the effects of interference on communication channels (Lucky, 1965). It can also be used to produce an inverse model of an unknown plant, which is in some sense a best fit to the reciprocal of the unknown plant transfer function. The delay of Δ samples in Figure 1 is to allow for the delay, or propagation time, through the plant and the adaptation associated with the inverse modeling. Including such a delay generally results in a much lower value for the minimum mean-square-error and causes the output of the converged adaptive system, y_k , to approximate the input, u_k , with a delay of Δ .

The mathematical description for this model can be expressed in discrete system form as follows. The input vector of the inverse model, Y , which represents a window of L values associated with the k th sample, is the input to the delayed inverse model (assuming n_k is zero). It can be expressed as:

$$\begin{aligned} \text{(single input) } \quad Y_k &= [y_k \ y_{k-1} \ \dots \ y_{k-L+1}]^T; \\ \text{(multiple inputs) } \quad Y_k &= [y_{1k} \ y_{2k} \ \dots \ y_{Lk}]^T \end{aligned} \quad (1)$$

for the single input case or for multiple inputs, respectively, and the weight vector, W_k , associated with the k th sample can be expressed as $W_k = [w_{1k} \ w_{2k} \ \dots \ w_{Lk}]^T$. The expression for w_{ik} ($i = 1, \dots, L$) will depend on the approach used. The elements of the input vector for the inverse model, Y_k , are considered to be L sequential sample inputs from a single input source, or L simultaneous inputs from L different input sources in the multiple inputs case. The elements, y_j (where $j = k, k-1, \dots, k-L+1$), of the single input vector Y , are set to zero if subscript j is equal to or less than zero (i.e., when $k < L$).

The output of the inverse model, V , for a linear combination of the input vector, Y_k , and weight vector, W_k , is written as

$$V_k = Y_k^T W_k = W_k^T Y_k \quad (2)$$

where, again, the subscript k is used as a time or sample number index. This function is the linear combiner or weighting function.

Thus the error associated with sample k from Figure 1 is

$$\epsilon_k = d_k - V_k = d_k - Y_k^T W_k = d_k - W_k^T Y_k \quad (3)$$

and the mean-square error is

$$\begin{aligned}\xi &= E[\epsilon_k^2] = E[d_k^2] + W_k^T E[Y_k Y_k^T] W_k - 2 E[d_k Y_k^T] W_k \\ &= E[d_k^2] + W_k^T R W_k - 2 P W_k\end{aligned}\quad (4)$$

where the input correlation matrix is $R = E[Y_k Y_k^T]$, and the cross correlation column vector is $P = E[d_k Y_k^T]$.

Our goal is to find the minimum mean-square error for the inverse model. Several adaptive methods, such as Newton's method and the steepest descent method, can be applied to adjust the weight vector in order to obtain the minimum mean-square-error. The disadvantage of these two methods is that they usually require off-line gradient estimation or repetition of data in order to compute ξ in equation (4) and its gradient, which reduces the computation speed for real time applications. By contrast, the LMS algorithm does not require an estimate of all the above terms, and computes the weighting function based only on instantaneous error. The details of the LMS algorithm are described in the next section.

3. The LMS Algorithm

In Newton's method or in the steepest descent method, the gradient of $\xi = E[\epsilon_k^2]$ is estimated by taking differences between short-term averages of ϵ_k^2 . However, for the LMS algorithm, ϵ_k^2 itself is used as an estimate of ξ_k without averaging.

The error in equation (3) is

$$\epsilon_k = d_k - Y_k^T W_k \quad (5)$$

and the estimated gradient of the mean-square-error can be obtained by differentiating equation (5)

$$\nabla_k = \frac{\partial \epsilon_k^2}{\partial W} = \begin{bmatrix} \frac{\partial \epsilon_k^2}{\partial w_1} \\ \vdots \\ \frac{\partial \epsilon_k^2}{\partial w_L} \end{bmatrix} = 2\epsilon_k \begin{bmatrix} \frac{\partial \epsilon_k}{\partial w_1} \\ \vdots \\ \frac{\partial \epsilon_k}{\partial w_L} \end{bmatrix} = -2\epsilon_k Y_k \quad (6)$$

For many practical adaptive system applications, the system transfer function is not fully determined and needs to be measured or estimated based on the sensed data. The slow adaptation (compared with Newton's method) used in the steepest descent method provides a filter process which reduces the effects of gradient measurement noise. Using the same slow adaptation as in the steepest descent method, the LMS algorithm's weighting update function is expressed as

$$w_{i,k+1} = w_{i,k} - \mu \nabla_k = w_{i,k} + 2 \mu \varepsilon_k Y_k \quad (i = 1, \dots, L) \quad (7)$$

where μ is the gain constant that regulates the speed and stability of the adaptation and has dimensions of reciprocal input power. Equations (5) and (7) are a form of the adaptive LMS algorithm proposed by Widrow et al. (1975). It has been pointed out by Widrow and Stearns (1985) that for the LMS algorithm, the mean of ε converges to zero as k approaches infinity provided that the following condition is satisfied: $0 < \mu < 1/\lambda_{\max}$, where λ_{\max} is the largest element in the diagonal eigenvalue matrix of R . Increasing the gain constant, μ , tends to reduce the number of iterations required for the LMS algorithm to reach its steady-state value, but also causes a corresponding increase in the average square error.

Only simple numerical operations (no squaring, averaging or differentiation) are performed in the LMS algorithm making this approach computationally more efficient than Newton's method or the steepest descent method. Since the change in the weight vector at each iteration is based on imperfect estimates, ε_k^2 , without averaging, the results obtained from the LMS algorithm may include higher frequency components (i.e., noise) and will not be a perfect match with the optimal (i.e., the case when $\varepsilon = 0$). Usually, the more one wants to attenuate the noise, the longer the computation time, but given the nature of the LMS algorithm, this is not a significant factor. The expected residual error can be reduced by using closed loop control.

4. Control Based On An Adaptive Inverse Model

Based on the inverse model described in section 2 with weights adapted by the LMS algorithm, we developed a delayed adaptive inverse model controller. The controller is designed to generate the approximate inverse of the plant given the desired and actual outputs. For a non-minimum phase system, some of the plant transfer function zeros are located in the right half of the s -plane or outside the unit circle in the z -plane for the discrete (or digital) case. In such cases, the exact inverse model (or the reciprocal transfer function) of the plant will have poles in the right half of the s -plane or poles outside the unit circle in the z -plane. Thus, the control input of a plant computed by using an exact inverse model will always continue to increase in magnitude with time and be unbounded. This will lead to an unstable system which will most likely drive the system into saturation. In order for such an inverse model to be stable, the impulse response with adaptation weights would need to be left-handed in time, having a non-zero input and output before the start time (Haykin, 1986). Including a delay lets the adaptive inverse model have a two-sided impulse response, thus solving the instability problem. A delayed adaptive inverse model (Widrow and Stearns, 1985) which can be used for control of a non- or unknown minimum phase plant, as shown in Figure 2, will give an approximated but stable system. This means that the stability of the controller can be assured regardless of whether or not the plant is minimum phase.

The delayed adaptive inverse model in the right half of Figure 2 is essentially the same as that shown earlier in Figure 1. The controller version on the left is a copy of the delayed adaptive inverse model. The weights are computed and updated on-line for real-time implementation of the

delayed adaptive inverse control algorithm. Although the plant transfer function in Figure 2 may be entirely unknown, certain information regarding the plant characteristics would be useful for selecting the number of weights and the length of delay. The number of weights should be proportional to the complexity of the plant; the higher the order of the plant, the larger the number of weights should be. The delay should approximate the propagation time from the input of the plant to the output of the delayed adaptive inverse model. A general rule of thumb is to set the delay, Δ , to be half of the number of weights. Once a convergent set of weights is obtained, the delay is increased or decreased for an optimal solution.

The approaches for updating the weight vector in the delayed adaptive inverse model can be explained using the schematic shown in Figure 3, where L is the number of weights used in the delayed adaptive inverse model, t_s is the sample period, t_i is the time needed per iteration, and t_c is the time needed for convergence of the weights for a given sample set of inputs and outputs of the plant, which is equal to the number of iterations to convergence multiplied by t_i .

One approach is to update all the weights only once at each sampling time, in which case t_s is equal to t_i . The number of sample periods used in this approach must of course be much greater than the number of weights in order to achieve convergence to the desired value. After a change in the plant's dynamic characteristics, a considerable number of sample periods (larger than the number of weights) may be required in order to reach convergence. The second approach is to iterate the weights using the LMS adaptation several times during each sampling time, in which case t_c is equal to or less than t_s . The first approach is applicable to very high sampling rate systems. The second approach is applicable when faster processors are available or for systems in which the rate of convergence is faster than the sampling rate. This convergence rate will be a function of the complexity of the plant. If the rate of convergence is slower than the sampling rate, one can use the maximum number of iterations that are possible within the sampling period. Of course, the permissible number of iterations within each sampling period is processor dependent.

5. Control of a Flexible Beam Unit: Simulation Results

The application of the delayed adaptive inverse method to a flexible beam unit with joint and structural flexibility was reported in Yang (1991). It was demonstrated that the delayed adaptive inverse method works for controlling systems with differentiable nonlinearities. In this paper, more recent results for the delayed adaptive inverse control of a flexible beam system will be presented. The system that we will consider here is a testbed located at the Armament Research, Development and Engineering Center (ARDEC) to simulate the motion control problems found on typical Army systems. A simplified schematic of the testbed is shown in Figure 4; the support and bearing structures are not shown for clarity. A linear model of the testbed developed by Bhat (1991) is provided in Figure 5. Elements of the block diagram model shown in this figure are described from left to right. The first block is the torque constant of the actuating motor with

$$K_t = 0.89 \text{ N-m/volt} . \quad (8)$$

The second block represents the motor dynamics described by

$$\frac{60}{s^2} \text{ radians/N-m} . \quad (9)$$

where s is the Laplace variable. The third block represents the compliance and damping of the shaft; it has the form $Bs+K$, where B is the damping term and K is the spring constant of the shaft. They are

$$B = 0.1663 \text{ N-m/rad-sec}^{-1}, \text{ and } K = 38 \text{ N-m/radians} . \quad (10)$$

The output of this block is the effective torque applied to the inertia wheel. The next block represents a state space model of the inertia wheel and the associated flexible beam dynamics. Two outputs of this block are the inertia wheel angular rotation and the beam's tip acceleration. The last block is a double integrator which will yield the tip displacement in radians. Bhat's state space model for the inertia wheel and the flexible beam was developed using finite element methods by modeling the beam as four identical beam elements. Figure 6 shows the inertia wheel and the beam element partitions modeling for finite element analysis. The beam element's mass and stiffness matrices are given by

$$K_{el} = \frac{EI}{l^3} \begin{bmatrix} 12 & 6l & -12 & 6l \\ 6l & 4l^2 & -6l & 2l^2 \\ -12 & -6l & 12 & -6l \\ 6l & 2l^2 & -6l & 4l^2 \end{bmatrix} \quad (11)$$

$$M_{el} = \frac{\rho A l}{420} \begin{bmatrix} 156 & 22l & 54 & -13l \\ 22l & 4l^2 & 13l & -3l^2 \\ 54 & 13l & 156 & -22l \\ -13l & -3l^2 & -22l & 4l^2 \end{bmatrix}, \quad (12)$$

where E , ρ , A , I , l are the Young's modulus, the density, the cross sectional area, the moment of inertia, and the element length of the beam, respectively. Each segment of the finite element of the beam satisfies the relationship (the terms in the equation are specified in Figure 6b.)

$$M_{el} \begin{bmatrix} \ddot{x}^{(1)} \\ \ddot{\theta}^{(1)} \\ \ddot{x}^{(2)} \\ \ddot{\theta}^{(2)} \end{bmatrix} + K_{el} \begin{bmatrix} x^{(1)} \\ \theta^{(1)} \\ x^{(2)} \\ \theta^{(2)} \end{bmatrix} = \begin{bmatrix} F^{(1)} \\ M^{(1)} \\ F^{(2)} \\ M^{(2)} \end{bmatrix}. \quad (13)$$

The root locus of this inertia wheel and beam model, shown in Figure 7, indicates that there are three right half plane zeros in the Laplace domain; this information confirms that the system is indeed non-minimum phase due to the non-collocatedness of actuation and sensing locations. A comparison of the frequency responses of the linear model of the flexible beam derived by Bhat (1991) from motor torque input to beam tip position with the experimental test results is shown in Figure 8. For the analysis proposed in this paper, reasonable agreement was observed over a range of frequencies. The performance objective to be achieved is that the tip of the beam should accurately track a desired path with specified boundary conditions on the motion. In this study, we will focus on the use of the accelerometer as the tip motion sensor in order to explore the limit of this readily available sensor. The desired acceleration profile will be used as the input command, $r(t)$, (i.e., the desired motion) of the controller in Figure 2. Since the output of the controller is a control torque, the adaptive inverse model is generating a characterization of the plant that can be described as an inertial effect, i.e., $\tau = I \ddot{\theta}$.

The desired displacement, velocity, and acceleration trajectories are generated by a fifth order Hermite polynomial (Forrest-Barlach and Babcock, 1986) to ensure that the desired motion of the tip satisfies the initial and final conditions as given in equation (14).

$$y_d(0) = \dot{y}_d(0) = \ddot{y}_d(0) = 0; \quad y_d(t_f) = y_{d1}; \quad \dot{y}_d(t_f) = \ddot{y}_d(t_f) = 0 \quad (14)$$

The desired tip displacement, velocity and acceleration can then be determined from the following set of equations which meet the constraints of equation (14)

$$\begin{aligned} y_d &= \left(\frac{6}{t_f^5} t^5 - \frac{15}{t_f^4} t^4 + \frac{10}{t_f^3} t^3 \right) y_{d1}(t_f) \\ \dot{y}_d &= \left(\frac{30}{t_f^5} t^4 - \frac{60}{t_f^4} t^3 + \frac{30}{t_f^3} t^2 \right) y_{d1}(t_f) \\ \ddot{y}_d &= \left(\frac{120}{t_f^5} t^3 - \frac{180}{t_f^4} t^2 + \frac{60}{t_f^3} t \right) y_{d1}(t_f) \end{aligned} \quad (15)$$

For the set of tests we considered, we arbitrarily picked a motion involving a 45 degree swing in one second followed by a one second hold. Therefore, in the above equation (15), t_f was set to 1.0 second, and $y_d(t_f)$ was set at 0.785 radians. The sampling period, t_s , is 2.5 msec corresponding to a sampling frequency of 400 Hz. The desired displacement, velocity, and acceleration profiles are plotted in Figures 9, 10, and 11, respectively. The desired beam tip motion, $r(t)$, entering the controller of Figure 2 was set equal to $\ddot{y}_d(t_f)$, i.e., only the acceleration profile was used. The number of weights used in the adaptive inverse model was 64, and the delay, Δ , was 29 sampling periods. A controller design and analysis software package, MatrixX³, is used for the simulation of the adaptive inverse controller.

The position response of the flexible beam for the specified acceleration profile is given in Figure 12 along with the desired position response. A delay is clearly visible. Both the commanded and the controlled tip motion are completed in less than one second as desired. The difference between the desired and actual displacement is shown in Figure 13. The large error which peaks near 0.6 seconds is due to the delay used in the delayed adaptive inverse controller. The delay (which is needed for causality) is a function of the propagation time through the plant plus the computational effect. The motion reversal period for this system that is characteristic of non-minimum phase systems is shown in Figure 14. This motion reversal period of 0.08 seconds, results in an effective delay in propagation through the plant dynamics. Without this delay in the delayed adaptive inverse controller, this adaptive approach will not converge. The computed input torque to the plant is shown in Figure 15. Note the three characteristics of the generated controller torque; first, there is no residual steady state torque; second, the duration of the input torque is greater than the duration of the desired motion due to the delay; and third, there is smooth transition at the beginning and end of the torque profile. The parameters of the delayed adaptive inverse

³ Available from Integrated Systems Inc. Santa Clara, CA.

controller are obtained via an iterative manual optimization process: the values were determined to work best with initial weights set at 0.015, μ set at 0.075, and with 20 iterations between each sampling period.

Although the parameters of the delayed adaptive inverse model are obtained via a manually derived iteration approach, the parameters have predictable effects on the system response. The gain constant, μ , determines the rate and stability of the adaptation. A large value of μ will shorten the time for convergence but it may also cause instability of the controlled system. Effects of μ on the system response are shown for three values of μ in Figure 16. It is seen from this figure that the smaller the μ , the larger the steady state error will be, since the convergence of the weights are slower for small values of μ . As expected, a large value of μ made the system response unstable. In this case, this occurred when μ was increased to 0.076.

In Figure 17, three responses are plotted for various values of the initial weights. The steady state error increased for larger values of the initial weights. In our simulations, the system response became unstable for initial weights greater than 0.0156 for this system. Therefore, the initial weights should be chosen iteratively to optimize the system response so that the control error, ϵ_k , (see Figure 1) is minimized within the iterations performed during each sampling time periods.

As discussed earlier, the number of the delay periods used in the delayed adaptive inverse method reflects the propagation time from input to output of the plant; and the general rule of thumb is to set the number of delay periods to be half of the number of weights. After iterative simulations, a delay of 29 sampling periods (in our case, the delay equals to 29×0.0025 seconds or 0.0725 seconds) was found to be optimal for our 64 weight FIR based controller. As shown in Figure 18, the larger the delay, the larger the steady state error of the system response. The system became unstable when the delay was reduced to 28 sampling periods. This result indicates that the propagation time of the flexible beam system must be greater than 28 sampling periods.

A residual error remains after the one second settling time (see Figure 13). This is because the adaptive inverse method is supposed to only yield an approximation of the inverse plant dynamics. A feedback control loop can be incorporated into the delayed adaptive inverse approach as shown in Figure 19 in order to remove the residual error in the system response. A linear feedback controller could be used for this purpose since the amount of residual error is small in magnitude. To illustrate this point, a simple proportional gain was used as a feedback controller to eliminate the error ripples of Figure 13. The position response is plotted in Figure 20. The error between this response and the desired value is shown in Figure 21. Although it is not clear from the figure, the steady state error has been reduced by 64%.

We designed an H_∞ based loopshaping controller using the methodology described in Doyle et al. (1992) for controlling the tip motion of our flexible beam (i.e., the plant described by Figure 5 and equations (8) through (13)). We used the Robust Controller Toolbox in MatrixX for the controller design. For comparison purpose, no uncertainty was considered in designing the controller, i.e., only the nominal performance was considered. In an attempt to obtain the position response shown in Figure 9, an initial design specification for determining the ideal complementary sensitivity transfer function for loopshaping was formulated as a settling time of one second with critical damping for a step input. However, an optimal controller was not obtained since this specification was too stringent and not achievable. Therefore, the design specification was relaxed to a settling time of one second with an overshoot of less than 20% of the steady state value for a step input. The resulting H_∞ based loopshaping controller is fourteenth order, as a result of the order of the system (which is twelfth order) and the order of the weighting function (which is

second order). The step response of the system with the H_∞ based loopshaping controller is shown in Figure 22. The design specification was met, but an undesirable oscillation appeared which lasted for up to six seconds. The motion response of the flexible beam tip using this H_∞ based loopshaping controller with the desired input (fifth order polynomial) as shown in Figure 9 (which represents the desired displacement profile) is given in Figure 23. Comparison of this response with the response of the system with an adaptive inverse feedforward combined with the linear feedback based controllers as shown in Figure 20 (and repeated in Figure 23 for comparison purpose) shows that the settling time has grown to 3 seconds. It would seem that the adaptive inverse controller with a proportional controller results in superior performance. However, it is important to note that the superiority of the delayed adaptive inverse controllers over H_∞ based loopshaping controllers is only demonstrated for nominal performance and not for robust performance. In fact, in the presence of measurement noise (we considered both 5% and 10% of the maximum torque input applied additively at the plant input), the H_∞ based loopshaping controller yields a stable response (similar to the response shown in Figure 23) while the delayed adaptive inverse controller results in an unstable system response. It is therefore imperative that high quality measurements be available when using a delayed adaptive controller.

6. Conclusion

The work described here presents the development of a delayed adaptive inverse method which may facilitate the closed-loop control of a flexible beam. An adaptation of a linear combiner was achieved by using the LMS algorithm to find a best fit to the reciprocal of a given plant transfer function. A copy of this delayed adaptive inverse model was used to compute the necessary torques required to control the non-minimum phase system as shown in Figure 2.

The simulation results for the system show that the applied torque is bounded and that one can achieve reasonable nominal performance as long as the sensing is of high quality. Comparison of the two system responses, one with adaptive inverse feedforward and linear feedback controllers, and one with H_∞ based loopshaping controller, shows that the first system has better response characteristics. Although the delayed adaptive inverse method works for nonlinear systems, in this preliminary study a linear system was used to demonstrate the performance of the delayed adaptive inverse controller so that the result could be compared with that of an H_∞ based loopshaping controller, which is applicable to linear systems only. Both the delayed adaptive inverse controllers and the H_∞ based loopshaping controllers ought to be able to handle model uncertainty. This will be a subject for future investigation.

Since there were no squaring, averaging or differentiating operations in the LMS algorithm (the main computational element of this approach), the delayed adaptive inverse method proposed here is computationally efficient. The on-line approach for updating the weights described in section 4 updates all the weights iteratively 20 times per sampling period (a not unreasonable expectation for a high performance system, which in our case iterates 20 times every 0.0025 seconds).

The delayed adaptive inverse method uses only knowledge of the desired values for the output (specified as an input $r(t)$ in Figure 2) and the actual plant output measurements in order to determine the control input, $u(t)$, to the plant. This method can be applied equally well to linear systems, systems with differentiable nonlinearities, or even unknown plants as long as a control input can be applied and the plant output can be measured. The question of handling systems with non-differentiable nonlinearities such as backlash, hysteresis, friction, etc. is still an open one and needs to be addressed further.

REFERENCES

- Asada, H. and Ma, Z.-D., "Inverse Dynamics of Flexible Robots: Modeling and Recursive Computation Using Virtual Link Coordinate Systems" *Proceedings of the American Control Conference*, 1989, pp. 2352-2359.
- Bhat, S.P., "Technical Report for the Advanced Weapons Test Bed" ISI report no. 7834-003, Integrated Systems Inc., June, 1991.
- Book, W.J., Maizza-Neto, O., and Whitney, D. E., "Feedback Control of Two Beam, Two Joint Systems with Distributed Flexibility" *Journal of Dynamic Systems, Measurement, and Control*, vol. 97, no. 4, 1975, pp. 424-431.
- Cannon, R.H. and Schmitz, E., "Initial Experiments on the End-Point Control of a Flexible One-Link Robot" *International Journal of Robotics Research*, vol. 3, no. 3, 1984, pp. 62-75.
- Doyle, J.C., Francis, B.A., and Tannenbaum A. R., *Feedback Control Theory*, Macmillan Publishing Company, New York, 1992.
- Ficola, A., Marino, R., and Nicosia, S., "A Singular Perturbation Approach to the Dynamic Control of Elastic Robots" *Proceedings of the 21st Allerton Conference on Communication, Control, and Computing*, 1983.
- Forrest-Barlach, M.G. and Babcock, S.M., "Inverse Dynamics Position Control of a Compliant Manipulator" *Proceedings of the IEEE International Conference on Robotics and Automation*, 1986, pp. 196-205.
- Ghorbel, F., Hung, J.Y., and Spong, M. W., "Adaptive Control of Flexible Joint Manipulators" *Proceedings of the IEEE International Conference on Robotics and Automation*, 1989, pp. 1188-1193.
- Haykin, S., *Adaptive Filter Theory* Prentice-Hall, Inc. New Jersey, 1986.
- Khorasani, K.M. and Spong, M.W., "Invariant Manifolds and Their Application to Robot Manipulators with Flexible Joints" *Proceedings of the IEEE International Conference on Robotics and Automation*, 1985, pp. 978-983.
- Lee, H. and Castelazo, I.A., "Nonlinear Feedback Control of a Flexible Robot Arm" *ASME Winter Annual Meeting DSC-Vol.6*, 1987, pp. 307-314.
- Lucky, R.W., "Automatic Equalization for Digital Communication" *Bell Syst. Tech. J.*, vol. 44, 1965, pp. 547-588.
- Meldrum, D.R. and Balas, M.J., "Direct Adaptive Control of Flexible Remote Manipulator Arm" *ASME 1985 Winter Annual Meeting, PED*, vol. 15, 1985, pp. 115-119.
- Nelson, W.L. and Mitra, D., "Load Estimation and Load-Adaptive Optimal Control for a Flexible Robot Arm" *Proceedings of the IEEE International Conference on Robotics and Automation*, 1986, pp. 206-211.

- Nicosia, S., Tomei, P., and Tornambe, A., "Feedback Control of Elastic Robots by Pseudo-Linearization Techniques" *Proceedings of the Conference on Decision and Control*, 1986, pp. 397-402.
- Schmitz, E., "Modeling and Control of a Planar Manipulator with an Elastic Forearm" *Proceedings of the IEEE International Conference on Robotics and Automation*, 1989, pp. 894-899.
- Siciliano, B. and Book, W.J., "A Singular Perturbation Approach to Control of Lightweight Flexible Manipulators" *International Journal of Robotics Research*, vol. 7, no. 4, 1988, pp. 79-90.
- Sorensen, B., Donath, M., Yang G., and Starr, R., "The Minnesota Scanner: A Prototype Sensor for 3D Tracking of Moving Body Segments" *IEEE Journal of Robotics and Automation*, vol. 5, no. 4, 1989, pp. 499-509.
- Tomei, P., Nicosia, S., and Ficola, A., "An Approach to the Adaptive Control of Elastic at Joint Robots" *Proceedings of the IEEE International Conference on Robotics and Automation*, 1986, pp. 552-558.
- Tomizuka, M., "Zero Phase Error Tracking Algorithm for Digital Control" *Journal of Dynamic System, Measurement, and Control*, ASME, vol. 109, 1987, pp. 65-68.
- Tsujio, S., "A New Approach to Inverse Dynamics of Flexible Manipulator Arms" *Proceedings of the USA-Japan Symposium on Flexible Automation*, 1988, pp. 375-382.
- Widrow, B., et al., "Adaptive Noise Cancelling: Principles and Applications" *Proceedings of IEEE*, vol. 63, no. 12, 1975, pp. 1692-1716.
- Widrow, B. and Stearns, S.D., *Adaptive Signal Processing* Prentice-Hall, Inc., New Jersey, 1985.
- Yang, G.B. and Donath, M., "A Delayed Adaptive Inverse Method for Position Control of a Flexible Robot Manipulator" *Proceedings of the Japan-U.S.A. Symposium on Flexible Automation*, ISCIE, Kyoto, Japan, July, 1990, pp. 949-953.
- Yang, G.B., "A Delayed Adaptive Inverse Method for End Point Motion Control of Flexible Beams" Ph.D. Thesis, Department of Mechanical Engineering, University of Minnesota, 1991.
- Yurkovich, S., Pacheco, F.E., and Tzes, A.P., "On-Line Frequency Domain Information for Control of a Flexible-Link Robot with Varying Payload" *Proceedings of the IEEE International Conference on Robotics and Automation*, 1989, pp. 876-881.

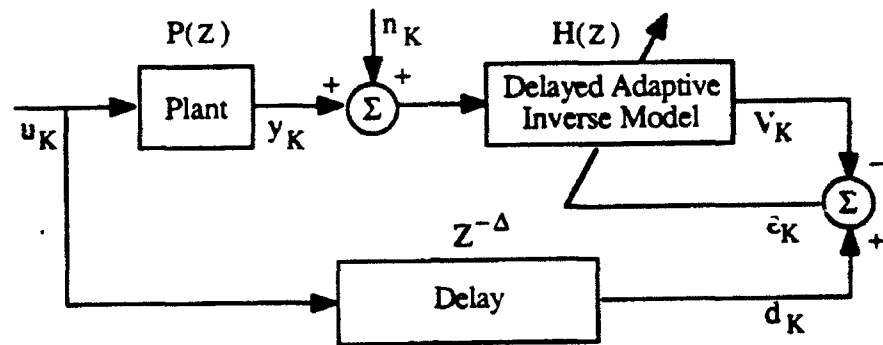


Figure 1. A delayed adaptive inverse model

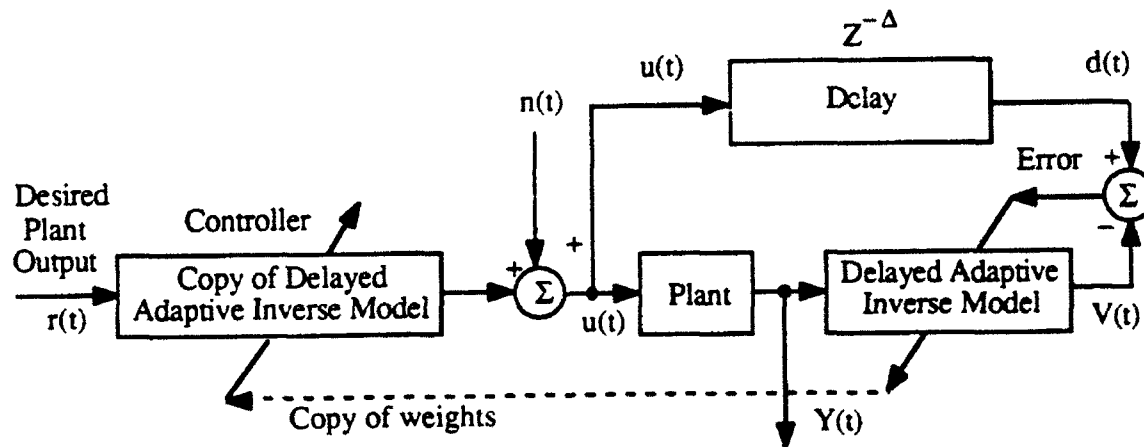


Figure 2. An adaptive inverse model based controller

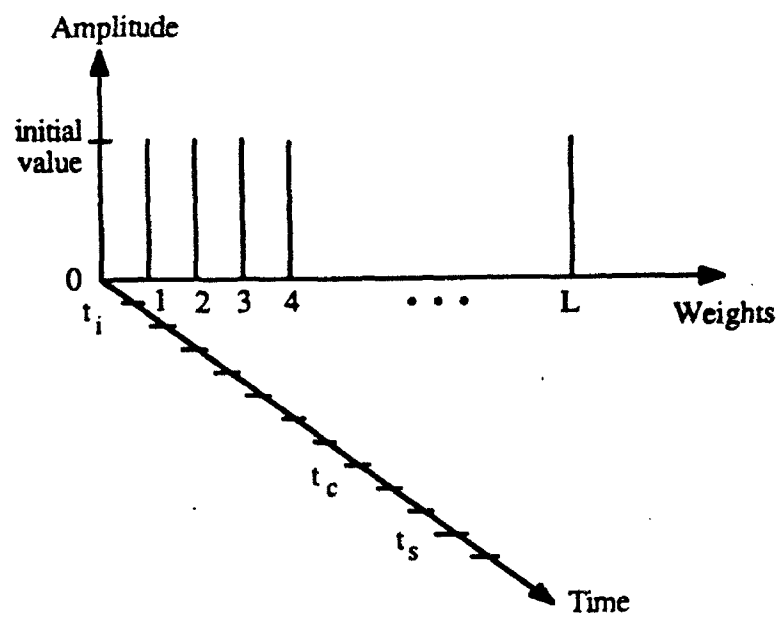


Figure 3. Schematic for weight updating

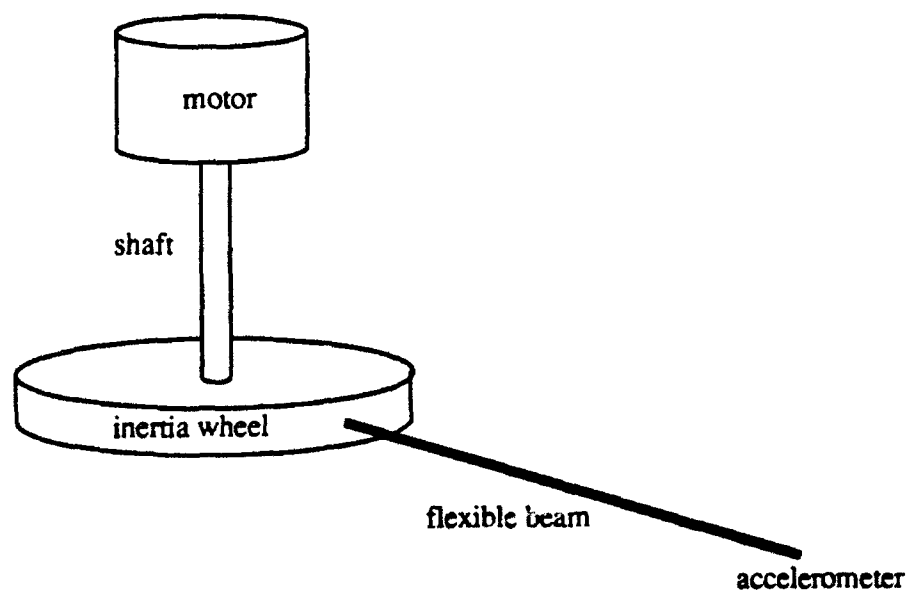


Figure 4. Schematic of the Picatinny flexible beam testbed

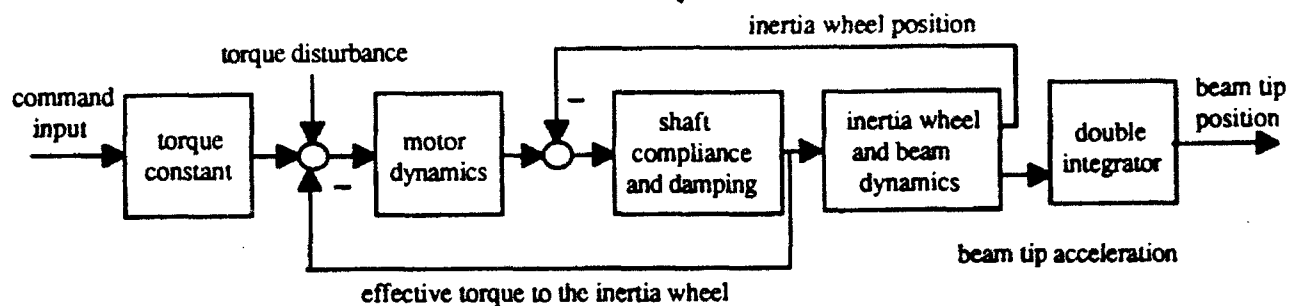
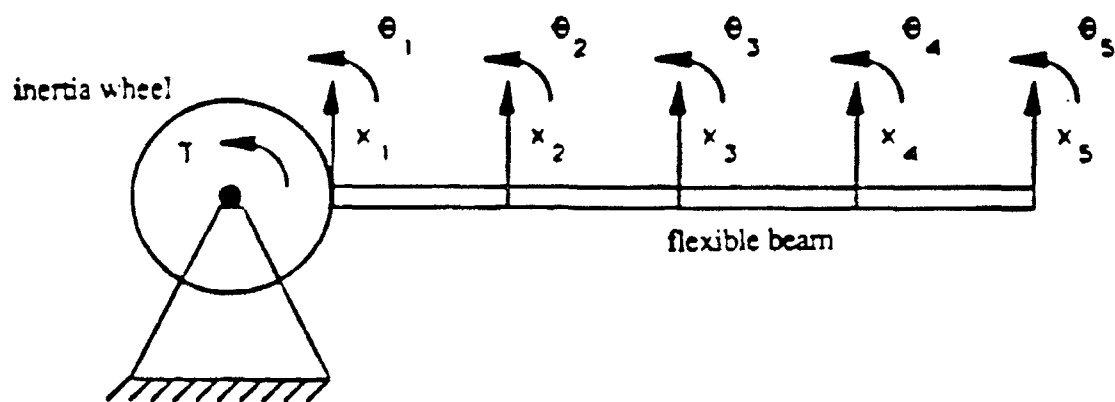
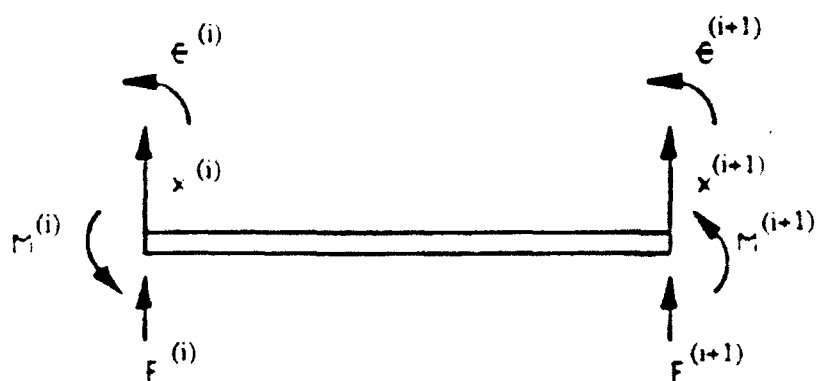


Figure 5. Linear model of the Picatinny flexible beam testbed



(a) Model of the inertia wheel and flexible beam



(b) an element of the flexible beam

Figure 6. Finite element model of inertia wheel and flexible beam
(from Bhat, 1991)

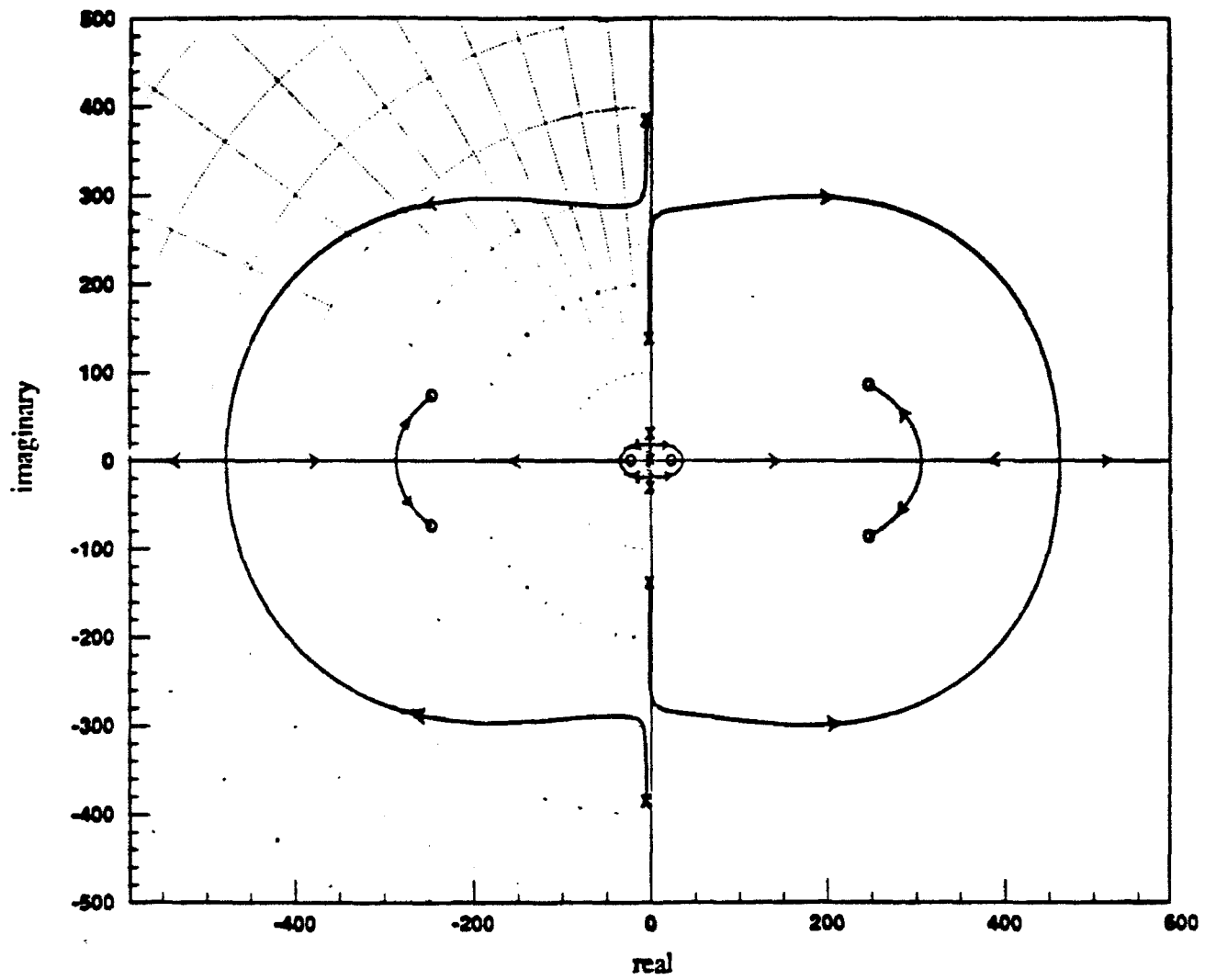


Figure 7. Root locus of the inertia wheel and beam model

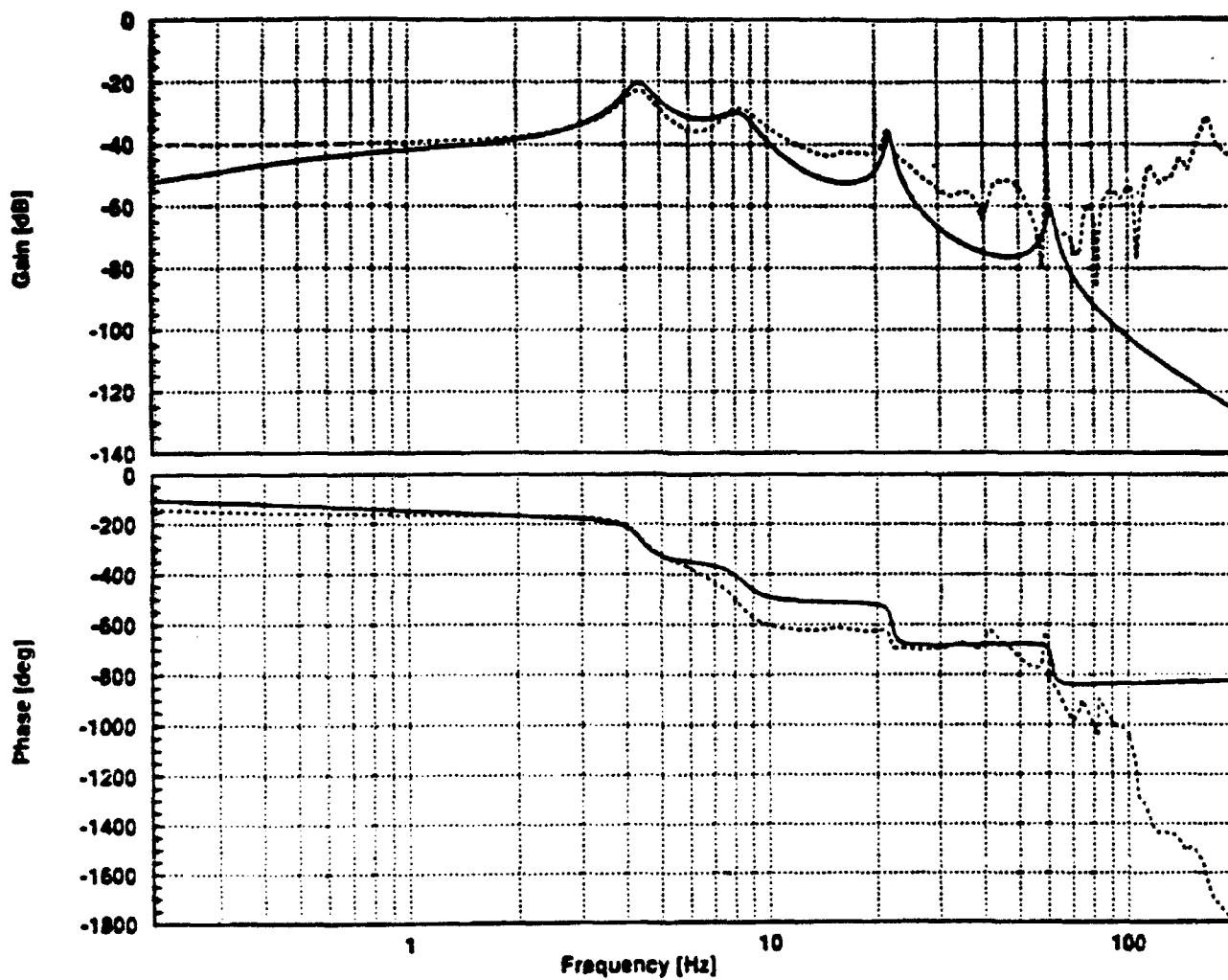


Figure 8. Analytical vs. measured frequency responses: torque to tip acceleration
analytical result (solid), measured result (dash)
(from Bhat, 1991)

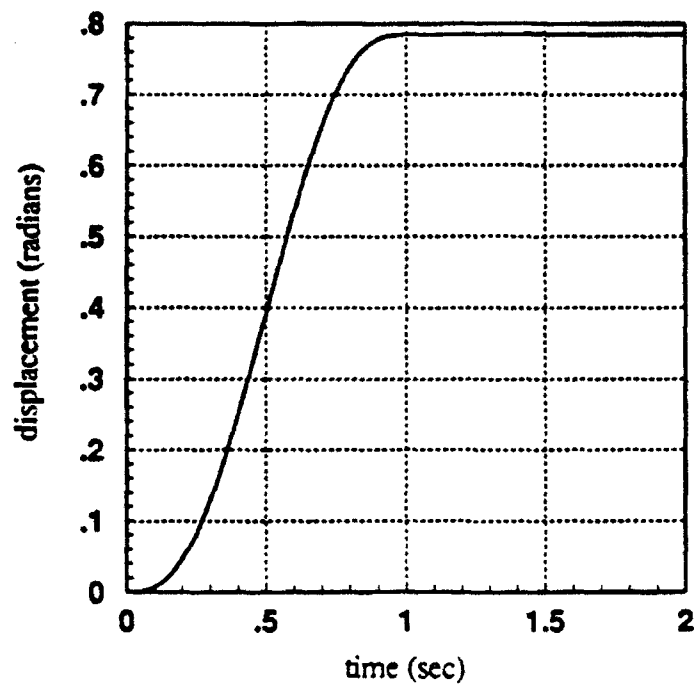


Figure 9. Desired displacement profile for the flexible beam

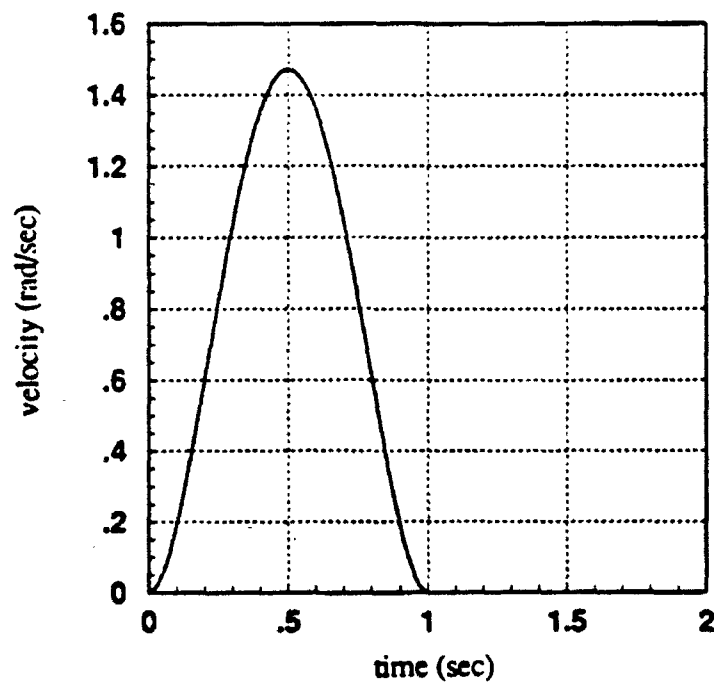


Figure 10. Desired velocity profile for the flexible beam

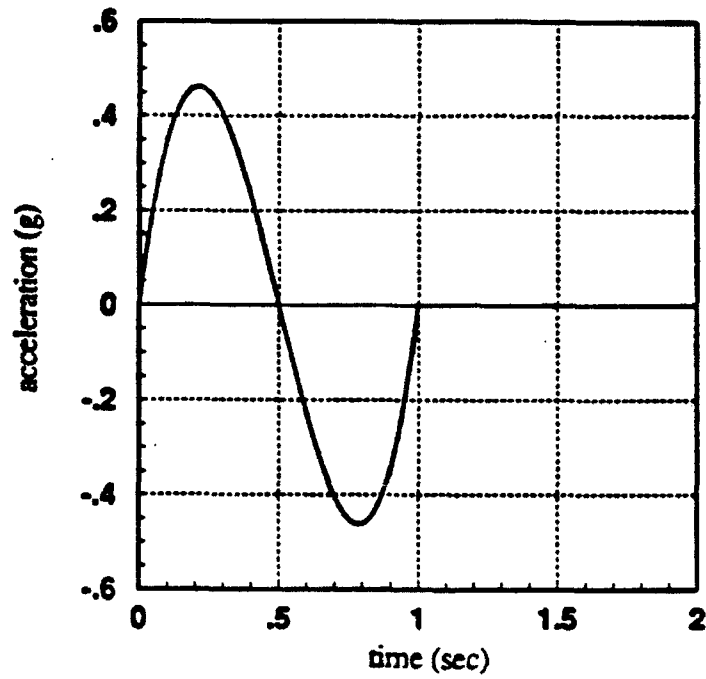


Figure 11. Desired acceleration profile for the flexible beam

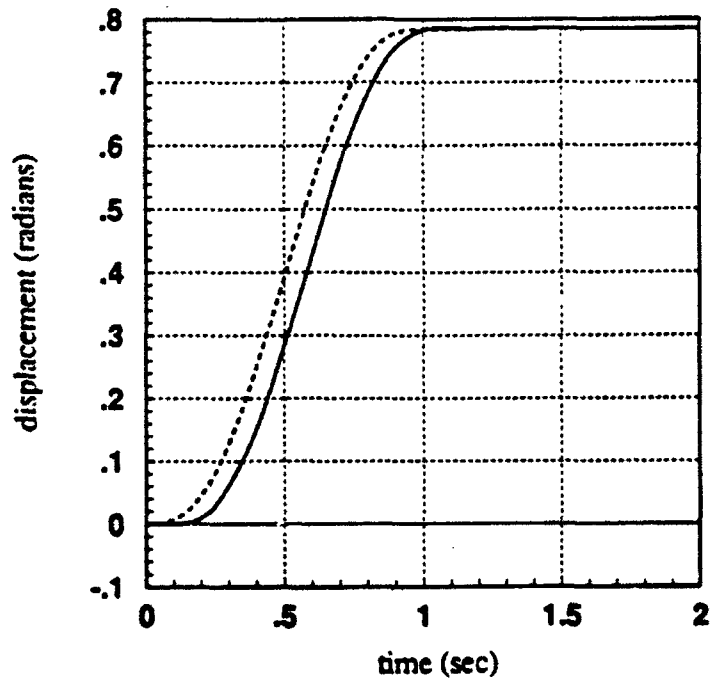


Figure 12. Simulated response of the flexible beam (solid) and desired response (dash)

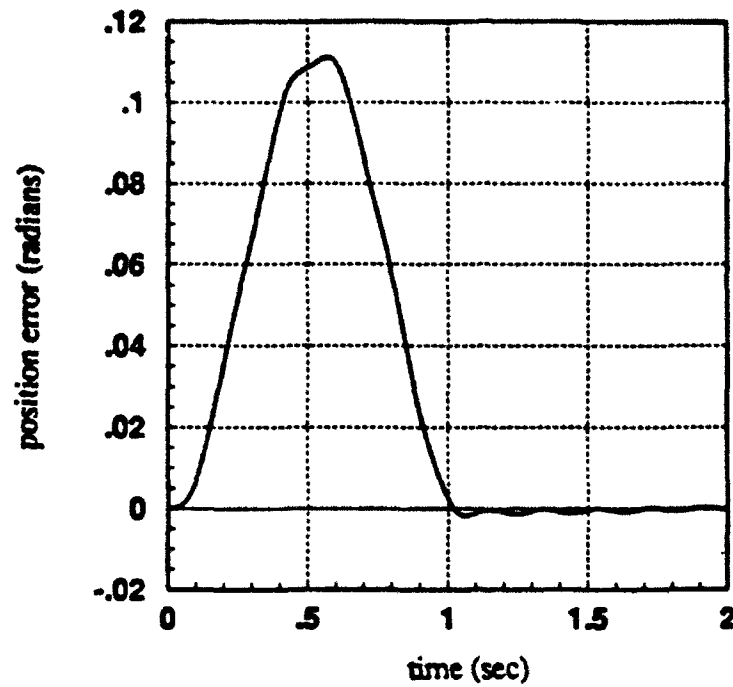


Figure 13. Tip position error for the delayed adaptive inverse method

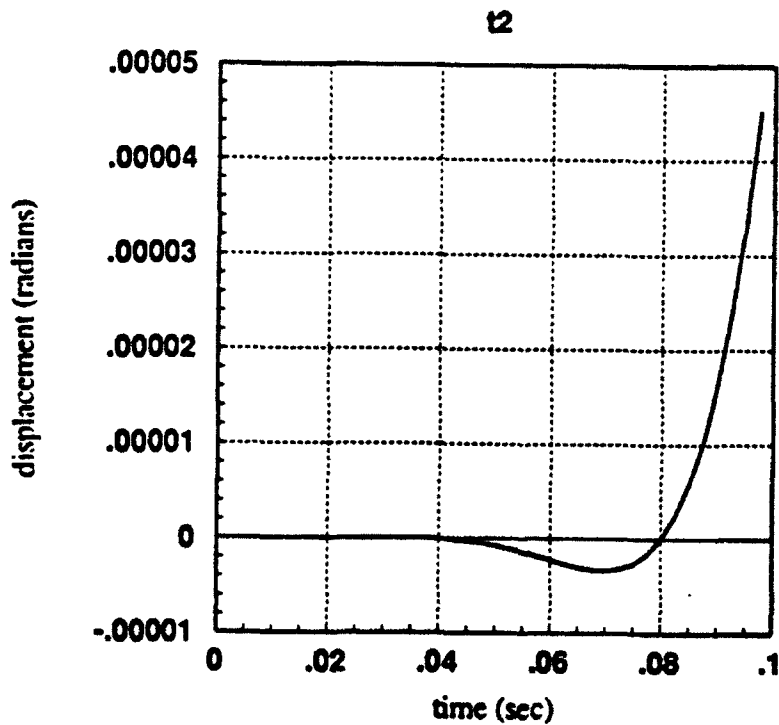


Figure 14. Close-up of Figure 12 showing motion reversal phenomenon

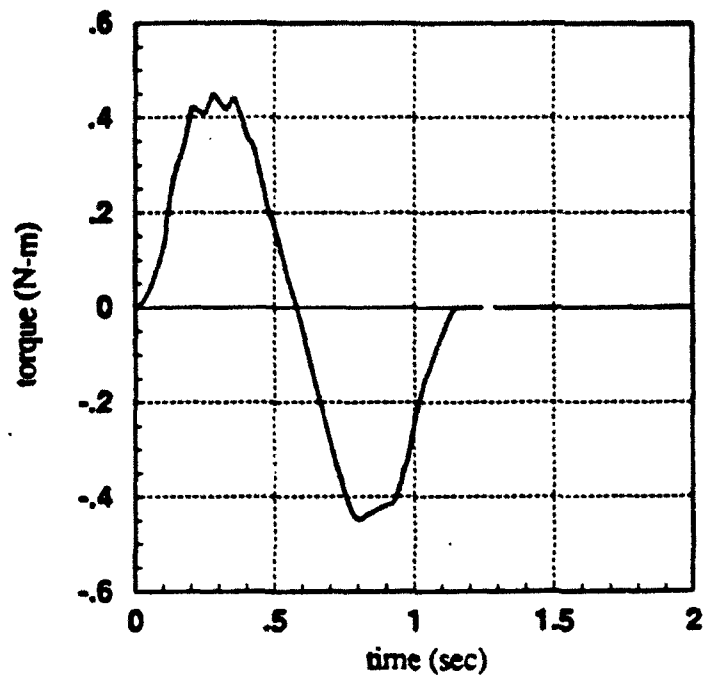


Figure 15. Input torque to the flexible beam

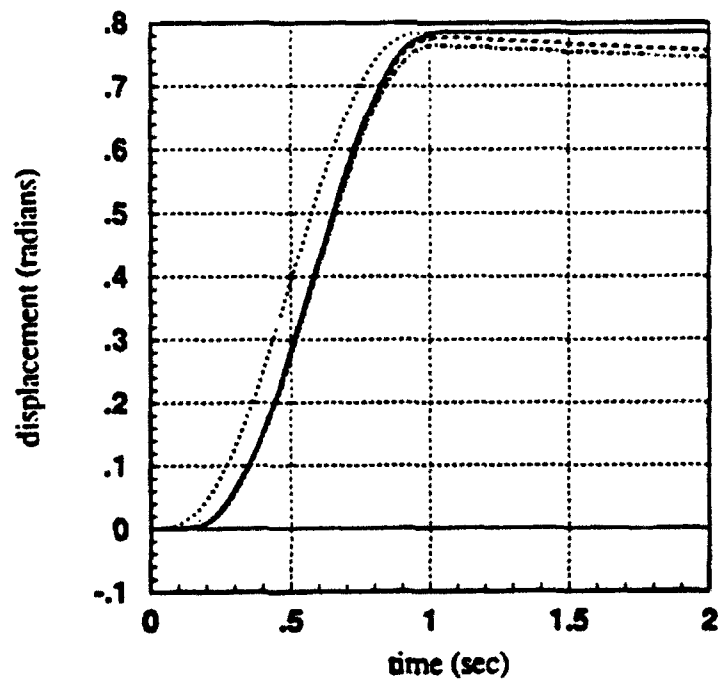


Figure 16. Effects of gain factor μ on the system response
0.075 (solid), 0.007 (dash), 0.0007 (dot-dash), desired (dot)

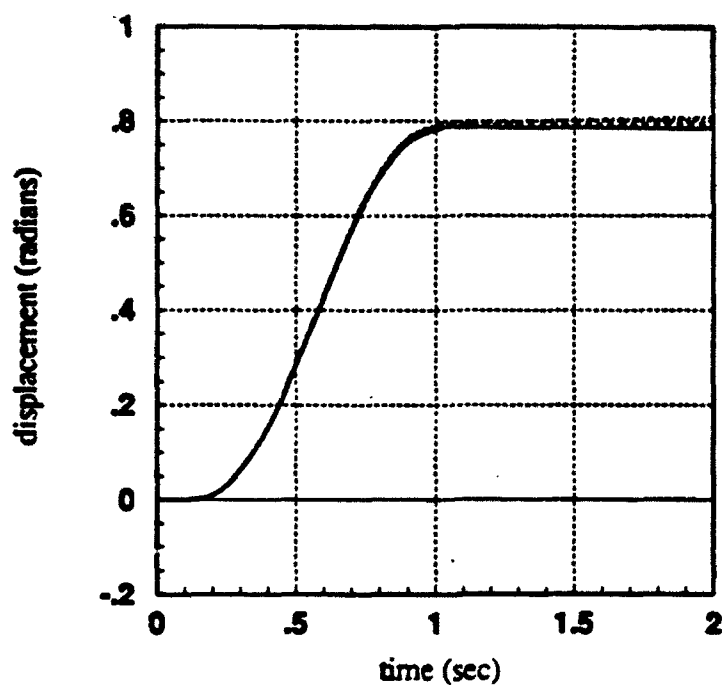


Figure 17. System responses for various values of the initial weights
0.015 (solid), 0.0152 (dash), 0.0154 (dot-dash)

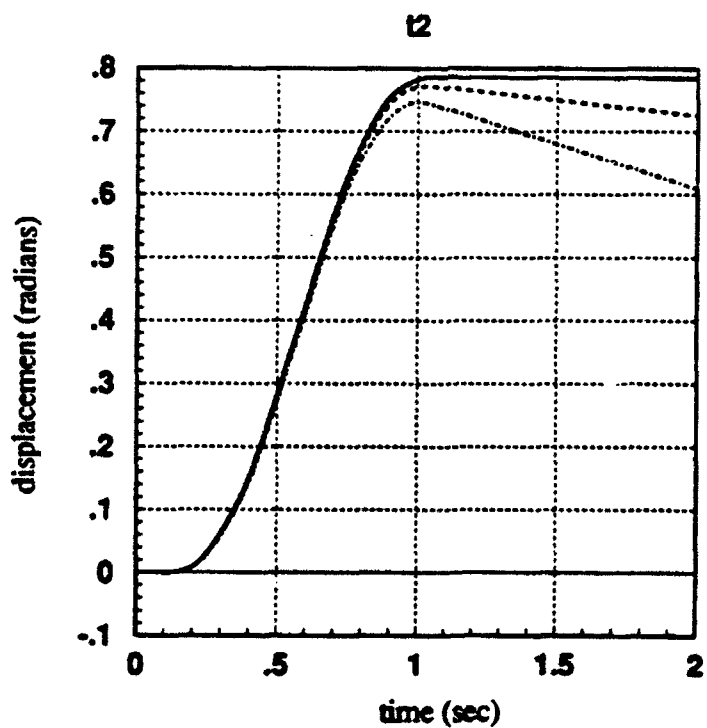


Figure 18. System responses for various number of delays
29 (solid), 30 (dash), 32 (dot-dash)

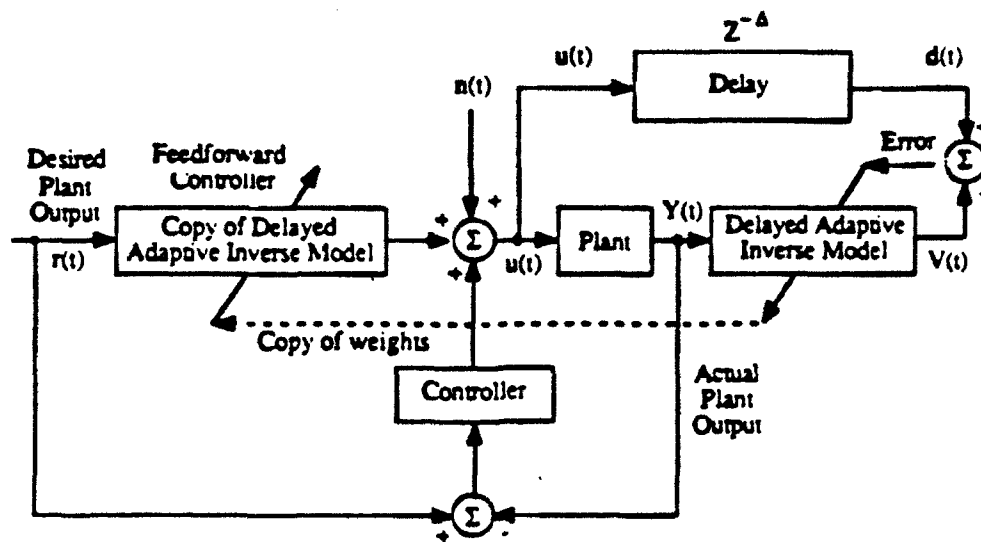


Figure 19. A closed loop control scheme based on the adaptive inverse method

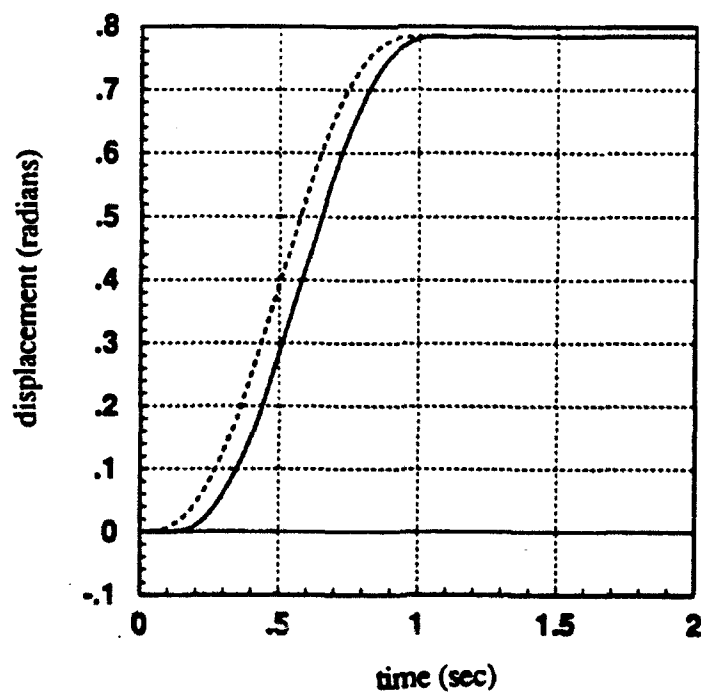


Figure 20. System response based on adaptive inverse and feedback with proportional gain
actual response (solid), desired response (dash)

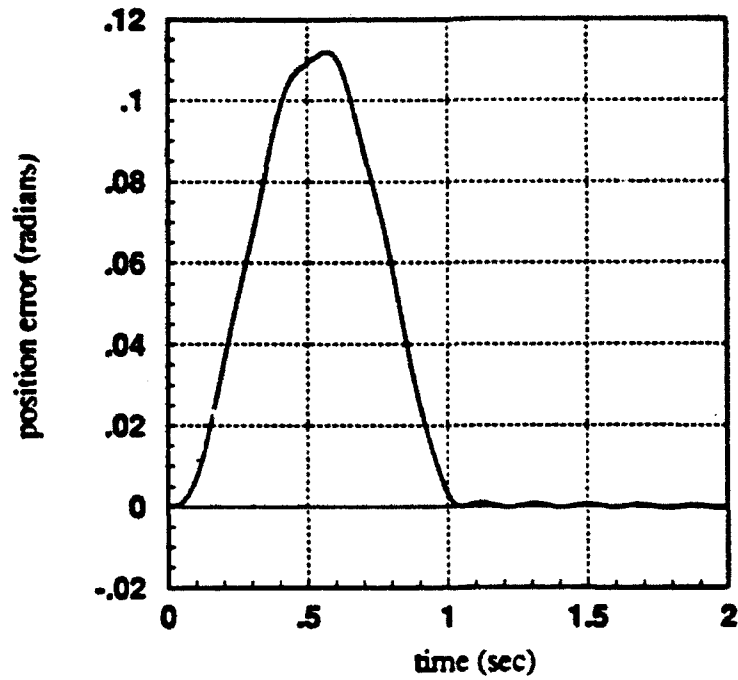


Figure 21. Tip position error for the adaptive inverse and feedback with proportional gain

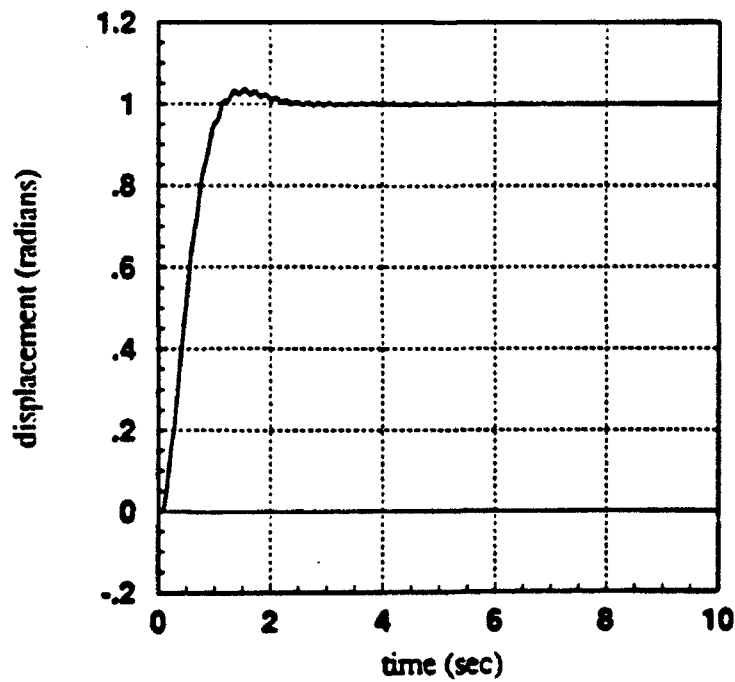


Figure 22. Step response of the tip position with an H_{∞} based loopshaping controller

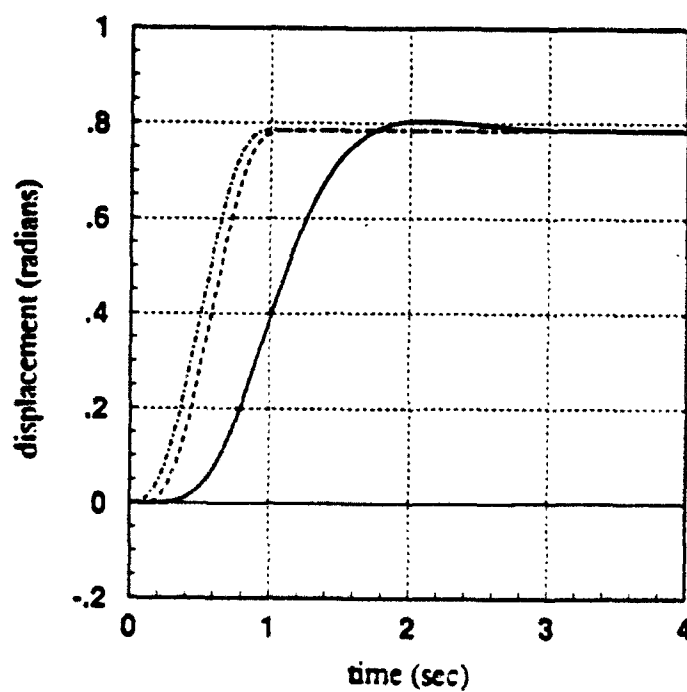


Figure 23. Response of tip position with an H_{∞} based loopshaping controller for an input shown in Figure 9
 actual response (solid), response of delayed inverse adaptive FIR method as in Figure 20 (dash), desired response (dot-dash)

Preliminary μ -synthesis Design for The ATB-1000

Dale Enns Daniel Bugajski
Honeywell Sysytems and Research Center
Minneapolis, Minnesota 55418

Allen Tannenbaum
Dept. of Electrical Engineering
University of Minnesota Minneapolis, Minnesota 55455

September, 1992

Abstract

A preliminary μ -synthesis controller for the Army's ATB-1000 test fixture is designed and analyzed. For comparison, two SISO controller designs are also described. The test fixture is patterned after the Apache helicopter's 30 mm gun and has tunable nonlinearities which may be representative not only of the nonlinearities of the gun, but of other mechanical systems as well. The models of the test fixture which were available at the time of the work are also described. The goal in pointing the gun is to reduce dispersions of fired gun rounds on targets. The resulting μ -synthesis design, when connected with a nonlinear simulation, exhibited limit-cycle behavior of unacceptable amplitude. The unacceptable performance is due to the nonlinearities and, in future work, would be improved upon by frequency domain trade-offs during the synthesis step.

1 Introduction

For proper overall functioning of most of the Army's weapons systems, specific subsystems demand high precision control. For example, a guided munition system may be fitted with laser systems for ranging and/or targeting. Both of the laser subsystems call for accurate pointing control systems. These are in addition to the high performance guidance and control of the munition itself. Tank and gun systems require stabilized platforms from which rapid firing and re-targeting occur. Stabilized platforms are also necessary

for antenna systems and video camera systems which are envisioned in future battlefield scenarios. Oftentimes, the accuracy of these control laws is limited by the mechanical system itself, for example, dead zones in gear drives, or friction in bearings.

The Army Research Office has built a laboratory fixture to study control laws for problems that are dominated by "hard" nonlinearities. Example nonlinearities in this group are saturations, static friction effects, and gear backlash. The fixture, the ATB-1000, is patterned after the Apache helicopter gun, and has built-in tunable nonlinearities. It is ideal for studying problems in application of linear and nonlinear control law designs.

This paper offers three potential linear control designs for the ATB-1000. Section 2 briefly discusses the objectives for the design, and Section 3 describes the models available for design and analysis. Section 4 discusses the three different designs, and Section 5 contains some nonlinear simulation and linear analyses for one of the designs. Finally Section 6 summarizes the work.

2 Requirements

The ATB-1000 is a test fixture patterned after the Apache's 30 mm gun. The basic goal for this weapon is to reduce dispersion of its rounds on targets. So the objective for the ATB-1000 is to minimize the barrel pointing angle deviation from a commanded value in the presence of platform motion (simulated with disk motion), gun firing-induced transients (simulated with a solenoid), and mechanism nonlinearities (simulated with adjustable backlash and friction). The laboratory fixture (see Figure 1) is outfitted with a laser arm to accurately measure the barrel tip position and hence experimentally determine performance. There are also disturbance levels and ranges of parametric nonlinearity adjustments that are part of the requirements.

3 Models

In practice, the development of a successful control system design is highly dependent on obtaining representative models of the system to be controlled. The models are a direct input to the control law synthesis and analysis steps in the development process. Models and modeling data come in many different forms, and different types of models are used for different purposes. Two distinct models of ATB-1000 test fixture were examined as part of this preliminary control design effort. These two models will be described and compared in this section. Some discrepancies between these two models have been identified and it will be necessary to resolve them for future studies.

During the summer of 1991, modeling data was received and analyzed from the Army Research Office. The modeling material consisted of MatrixX block diagrams, tables containing definitions, scale factors, sign conventions, units, signal size information, and linear

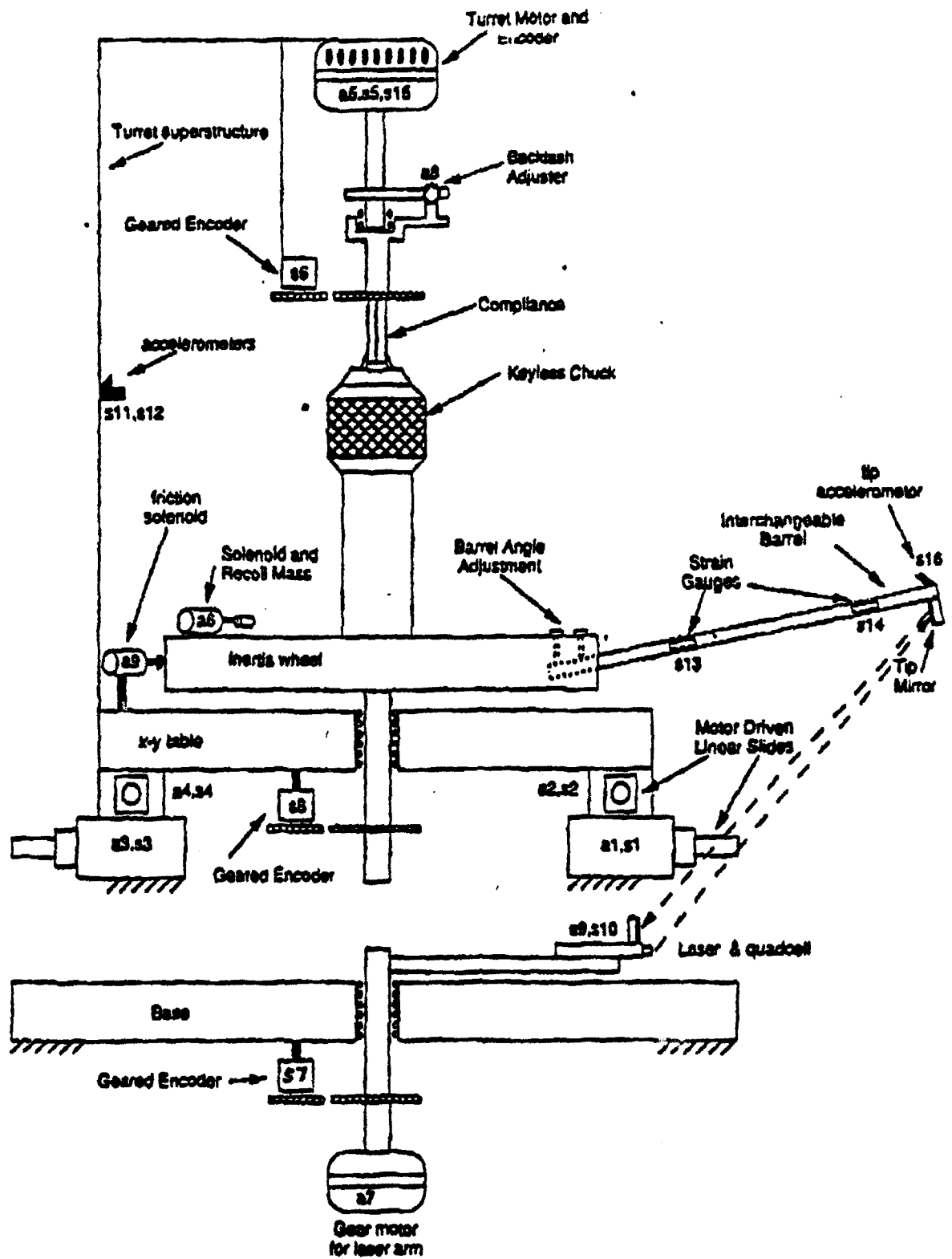


Figure 1. Schematic of ATB-1000 test fixture.

models for the truncated finite element model of the barrel (with 8 states) and a 27th order identification model. This section will refer to an analytical model and an identification model. The analytical model is based on the block diagrams, tabular data, first principles of dynamics, and includes four relay-type nonlinearities and two deadband-type nonlinearities. The identification model is a linear model whose inputs and outputs are a subset of those in the analytical model.

The MatrixX block diagrams and the tabular data were used to generate a linear model and a nonlinear simulation using Honeywell computer tools. The linear model treats the deadband as a unity operator which neglects backlash, and the relay as a zero operator which neglects friction. The linear model was then examined in terms of its poles, transfer function zeros for certain inputs and outputs, and time and frequency responses. There are degrees-of-freedom for the disk translation and rotation in a plane, motor rotation, inertia wheel rotation, laser arm motor rotation, laser arm rotation, and three elastic degrees of freedom for the gun barrel (simulated with a rod attached to the inertia wheel).

The linear open loop model consists of two physical systems: the disk, inertia wheel, and rod system, and the laser arm system. There is a motor associated with each system. There are five pairs of open loop poles at the origin (because friction is neglected) associated with the rigid body degrees-of-freedom. There are two pairs of complex poles associated with the compliances in both systems, and there are three pairs of complex poles associated with the gun barrel resonances with small damping ratios.

The plant transfer function between the control motor torque and the barrel pointing angle can be regarded as a double integrator (at frequencies below 10 rad/sec) with disturbances (from disk motion) and some high frequency elastic modes. The nonlinear simulation was executed with different test inputs to assess its behavior.

An identification (ID) model was obtained in a state space format with seven outputs, one input, and 27 states. The outputs are torque motor resolver, backlash resolver, disk velocity, quadcell output, strain gauge #1, strain gauge #2, and torque motor tachometer, the input is the control motor torque, and the 27 states are not physically defined but the linear ID model fits the data from the identification experiments. This model was compared to the analytical model in terms of poles, frequency response, and time histories.

The ID model shows more open loop damping e.g. $\zeta_{ID}=0.07$ versus $\zeta_{analytic}=0.01$ for the first elastic mode (near 31 rad/sec) and $\zeta_{ID}=0.15$ versus $\zeta_{analytic}=0.084$ for the shaft compliance mode (near 55 rad/sec) between torque motor and inertia wheel. The low frequency behavior of the ID model shows a slope of -1 on a Bode gain plot versus the slope of -2 in the analytical model, because friction is present in the identification experiment, but neglected in analytical model. In addition, the low frequency accuracy of the ID model is limited by the length of time used for the identification experiment. Thus the ID model is not close to the analytical model for frequencies below 10 rad/sec. Except for the poles and low frequency asymptote, the ID and analytical models agree for torque motor resolver, backlash resolver, and torque motor tachometer outputs. On the other

hand, the ID and analytical models for strain gauge #2 show 180 deg phase discrepancies, and the quadcell output does not show close agreement at any frequency.

For future design work, it will be necessary to resolve these discrepancies before closed loop testing can be performed. The ID model was utilized (despite these discrepancies) for demonstration of the μ -synthesis design methodology. Actually a balanced realization of the ID model was truncated to twelve states for the μ -synthesis design. The analytical model was utilized to develop alternate control laws with classical approaches. One of these classical alternates uses motor tachometer feedback, and the other has lead and notch compensation of the inertia wheel position. The next section discusses each of these three designs.

4 Control Law Design

In this section a preliminary design effort for the ATB-1000 test fixture will be described in detail. The design is incomplete, but adequately serves as a starting point for future work. To limit the scope of the preliminary effort, the "hard" nonlinearities were neglected for the control synthesis. However closed loop simulations were carried out where the nonlinearities were included. These preliminary simulations showed that the nonlinearities are significant and it will be necessary to include them in future designs. In this preliminary look at control law design, three design approaches were considered. Two approaches were SISO and one was multivariable μ -synthesis. The SISO designs are of interest because they correspond to minimal sensor requirements. The μ -synthesis approach is of interest because the nonlinearities are accounted for by treating them as bounded operators.

The control problem is to point the gun barrel in the face of disturbances. For the demonstration design presented here, the pointing was quantified in terms of the quadcell output and only the solenoid disturbance was included in the design objective. Model uncertainty was incorporated with a multiplicative perturbation at the torque motor location. Sensor noise was also included in the formal μ -synthesis problem statement. More detailed designs would incorporate frequency domain weighting transfer functions, which act as linear bounds for the effects of the six system nonlinearities. Requirements would also be defined and incorporated for actuator activity and physical limitations.

It was necessary to append a solenoid disturbance input (which simulates gun firing) to the ID model. This was done by selecting a constant gain matrix from the frequency response of the analytical model near the first elastic mode frequency. This is an approximation used for expediency during this preliminary design. In a more detailed design effort, the effect of the disturbance input on the equations of motion would be included more carefully into the state-space matrices for the interconnection structure used for μ -synthesis.

It is worth noting that the gun stabilization fixture is similar to a particular elastic

structure control problem which has received a large amount of attention in the control and modeling literature. In addition, experimental studies have been performed at various laboratories [3, 2]. The problem is that of rotating disks (at least two) that are connected with rods that are elastic in torsion. These studies motivated the first design.

Colocated SISO Design One of the SISO designs was for colocated feedback between the torque motor resolver and the motor torque. This choice was motivated by the knowledge that under certain assumptions regarding a lower bound for inherent structural damping, and sufficiently high bandwidth sensors, computers, and actuators, such a mechanical system can be robustly stabilized with colocated sensors and actuators even in the presence of some significant nonlinearities. When the sensor and actuator are not colocated, robust stabilization is, in general, more difficult to achieve due to limitations imposed by non-minimum phase aspects. [4, 5, 1]

The reduced order ID model was utilized to determine the feedback compensation. Recall that the transfer function has a $1/s$ shape below 10 rad/sec in the case of the ID model. Thus a pure gain can be selected to set the unit loop gain crossover at 10 rad/sec as a preliminary design choice. Higher frequency resonances are stabilized because of the colocation and the assumptions about inherent damping, sensors, and actuators. A higher crossover could be considered but this would require more accurate modeling of even higher frequency elastic behavior and tighter requirements on sensors and actuators. A pure gain feedback between motor position and motor torque would not be stabilizing if connected to the analytical model because it has a $1/s^2$ shape below 10 rad/sec as discussed above.

Noncolocated SISO Design The other SISO design was developed with the analytical model for noncolocated feedback between the inertia wheel encoder and motor torque. In this case a lead compensation element was employed to create a unit loop gain crossover at 10 rad/sec. In this case, some of the higher frequency resonances are destabilized by the noncolocated feedback. To prevent this destabilization, notch filters were included for the first elastic mode and the compliant mode between the motor and inertia wheel. This design approach is of interest (as compared to the colocated design) because the colocated motor position is not as closely related to the pointing angle as is the inertia wheel. This design also has value as a further comparison against the μ -synthesis design.

Mu-Synthesis Design The μ -synthesis design approach is multivariable and is cast in terms of the interconnection structure shown in Figure 2. There is a multiplicative perturbation at the torque motor location represented by Δ and the input v_1 and output

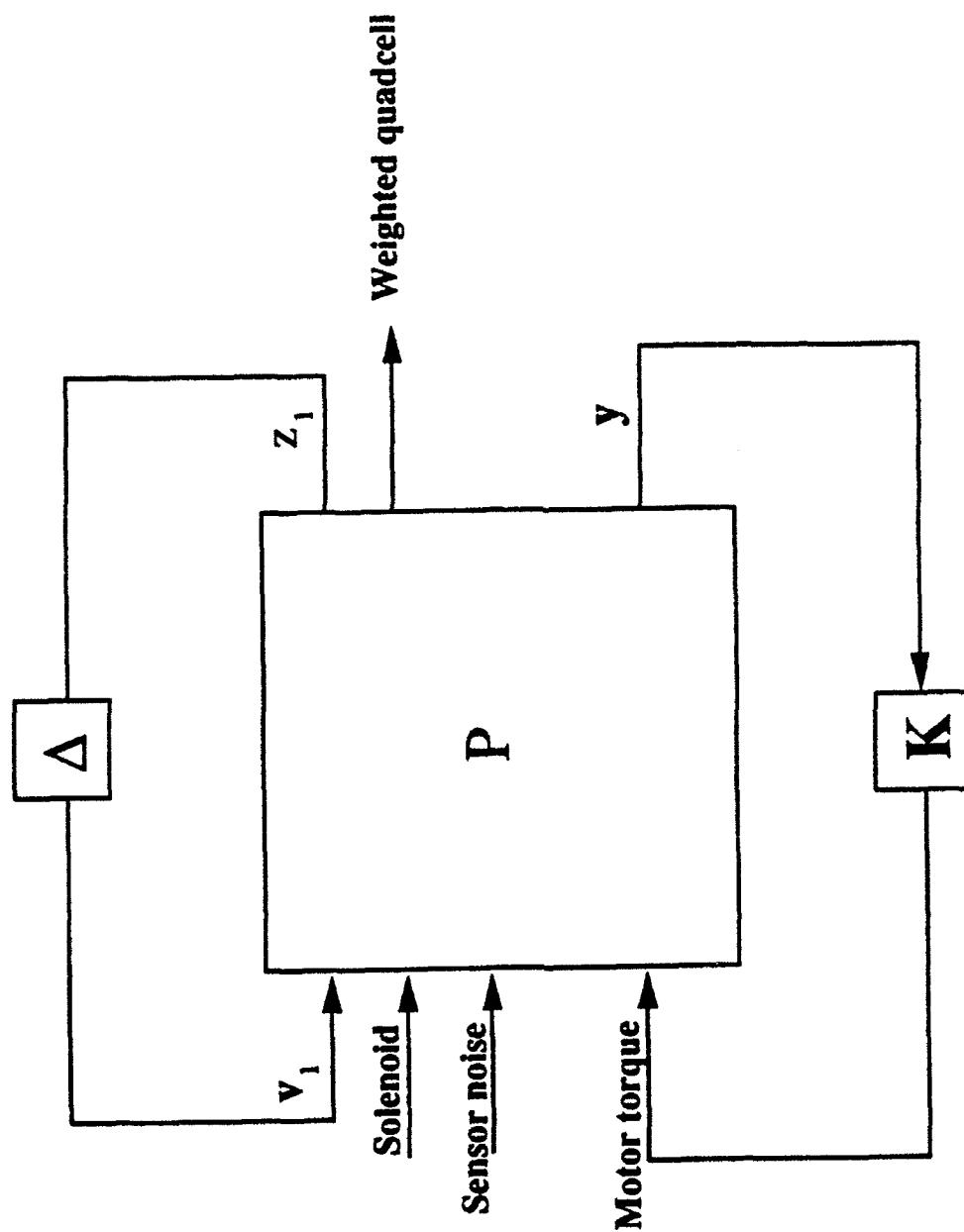


Figure 2. Block diagram of interconnection structure.

z_1 . There is a performance output called e_{quad} which is the quadcell output passed through a weighting function. The external inputs are sensor noise and the solenoid disturbance. There is also the torque motor input and the seven sensors to close the feedback loop with the compensator K .

The interconnection structure includes weighting transfer functions for uncertainty bounds, performance requirements, and disturbances. The uncertainty was modeled as a multiplicative perturbation and was bounded with a third order Butterworth filter having break frequencies at 20 and 300 rad/sec and a high frequency gain of 675. This can be interpreted as 20% model error below 20 rad/sec and 67,500% model error above 300 rad/sec. The pointing requirement is included by weighting the quadcell output with a low pass transfer function $360(s + 10)/(s^2 + 84s + 60^2)$. This has unit steady-state gain, so outputs of less than 1 volt would be acceptable. The seven sensor noises are weighted with the constant value of 0.01, so this corresponds to either volts or counts depending on the sensor. Finally, the solenoid disturbance is weighted with a low pass transfer function $0.3/(s + 10)$, so inputs of 0.03 volts are expected. The weightings were not carefully related to the hardware in this preliminary design demonstration. This relationship should be more carefully addressed to better account for known hardware characteristics. In particular a weighting for the disturbance would take into account the duty cycle of the solenoid. Additional inputs and outputs as well as weightings could be utilized to represent the nonlinearities which have not been accounted for in the preliminary design.

The state space solution to the H^∞ control synthesis problem was used to find a feedback compensator K . This compensator has as many states (18) as the interconnection structure and it was possible to reduce the compensator order by residualization to 16 states. The closed loop transfer function is denoted by M and connects the inputs: v_1 , solenoid, and sensor noise to the outputs: z_1 , and quadcell.

The next step in the μ -synthesis design was to introduce D -scales to properly account for the model uncertainty and performance variable response to external inputs. A constant D -scale=3 was employed because a dynamic D -scale was not deemed necessary in this preliminary design. The D -scale was incorporated by multiplying z_1 by 3 and dividing v_1 by 3 (i.e., $DM D^{-1}$) and a new interconnection structure P was established. The H^∞ problem was then re-solved for the compensator K and the iterations were terminated. Detailed analyses of this compensator appear in the following section.

5 Analyses

The μ -synthesis results are graphed in Figure 3. There are five plots of Bode magnitude versus frequency. The top curve is relatively flat because it is the maximum singular value of the closed loop interconnection structure ($\bar{\sigma}[M]$), and H^∞ optimization makes its peak value as small as possible. The next curve down is the structured singular value, $\mu[M]$,

musyn design

m.s, m.mu, m11.s, m22.s, m.sn.s

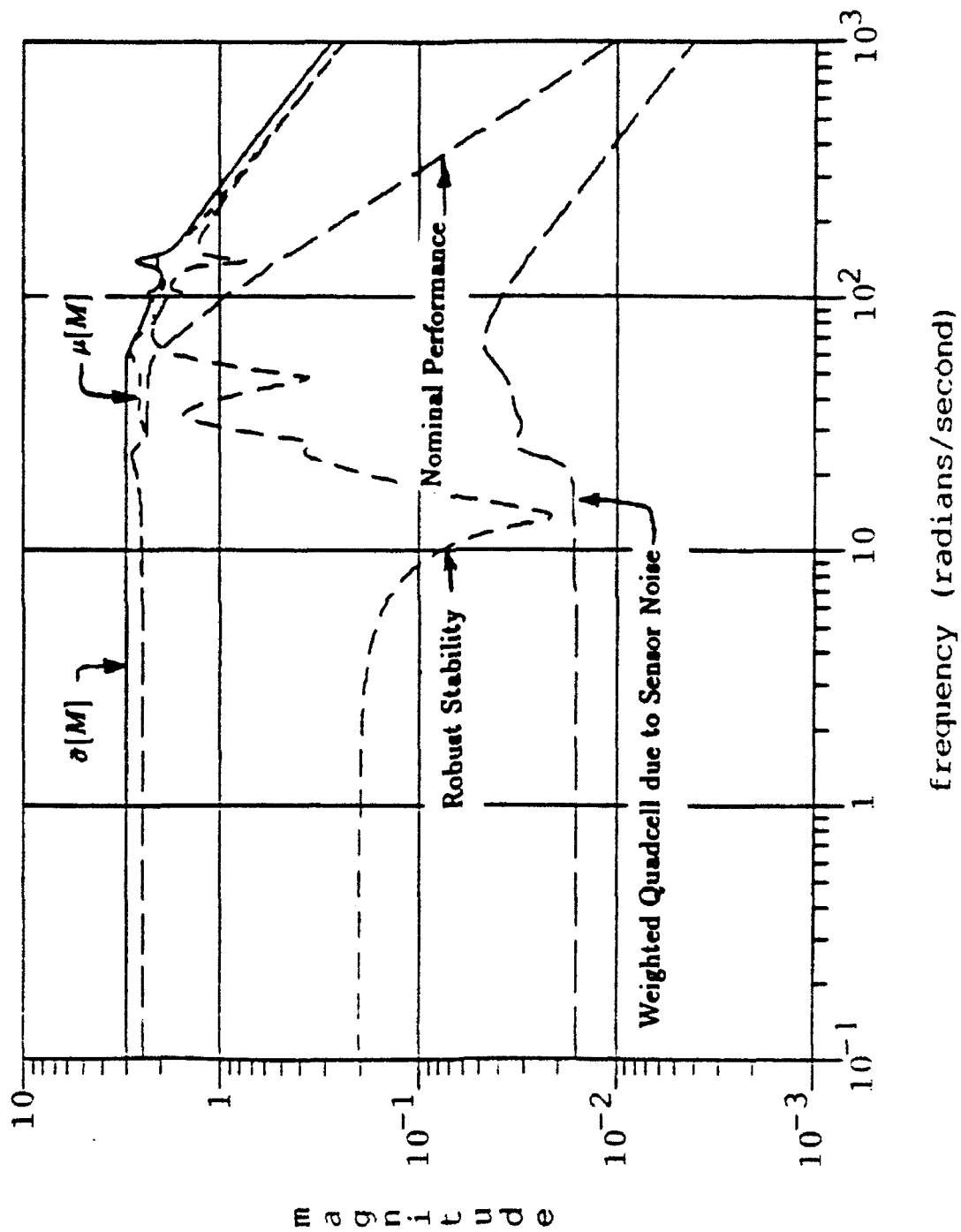


Figure 3. Graphical summary of mu-synthesis design.

and is necessarily less than or equal to the upper curve, since $\bar{\sigma}[M]$ is a theoretical upper bound for $\mu[M]$. There is a low frequency difference between the structured and maximum singular values, which indicates that performance improvements are possible by further D - K iteration and frequency dependent D -scales.

The next two curves in Figure 3 correspond to robust stability and nominal performance. Theoretically these curves are less than or equal to the structured singular value and this is consistent with the numerical results. The robust stability curve is relative to the defined multiplicative perturbation, and dominates $\mu[M]$ at higher frequencies. The robust stability curve can be further interpreted as the weighted complementary sensitivity, where the weighting is the bound for the multiplicative perturbation. The nominal performance curve is the maximum singular value of the transfer function matrix between the weighted quadcell and the external inputs including the weighted solenoid disturbance and sensor noise. This curve dominates $\mu[M]$ at low frequencies and can be further interpreted as the weighted sensitivity. The lowest curve in the figure corresponds to the weighted quadcell response due to sensor noise. This is more than an order of magnitude less than $\mu[M]$, so the quadcell/sensor noise path does not have much influence on the optimal design.

Further analyses of the μ -synthesis design were carried out to assess closed loop poles, input and output loop properties, and time response to solenoid disturbances. The closed loop poles indicated closed loop stability and damping improvements for the first elastic mode ($\zeta_{CL} = 0.12$ versus $\zeta_{OL} = 0.08$) and compliant mode ($\zeta_{CL} = 0.21$ versus $\zeta_{OL} = 0.14$). Gain and phase margins for the SISO loop transfer function at the torque motor actuator location were evaluated. The lowest frequency unit gain crossover occurs at 5.8 rad/sec with a phase margin of 81 degrees. The phase margins surrounding the first elastic mode frequency are larger than 43 degrees. All gain margins are larger than 7 db. These are considered good margins with respect to model uncertainty at the actuator location.

The linear closed loop system was simulated with the disturbance model used for the μ -synthesis design. This disturbance model is a constant gain matrix between the solenoid and the measurements including the quadcell output. Thus this model is only accurate near the first elastic mode frequency and is not accurate at low or high frequencies. The disturbance input was a 10 Hz sequence of 10 msec, 1 volt pulses. (See Figure 4a.) The quadcell output response is dominated by the compliant mode because the pulse frequency is close in proximity to the compliant mode frequency. The quadcell output during the 10 msec solenoid firing is not accurate, so disregarding these portions of the response, the quadcell output shows a residual oscillation near the compliant mode frequency with less than 3 volts peak-to-peak. (See Figure 4b.) This is not considered satisfactory performance and the interconnection structure should be further refined to improve the performance by making better tradeoffs with the weighting functions.

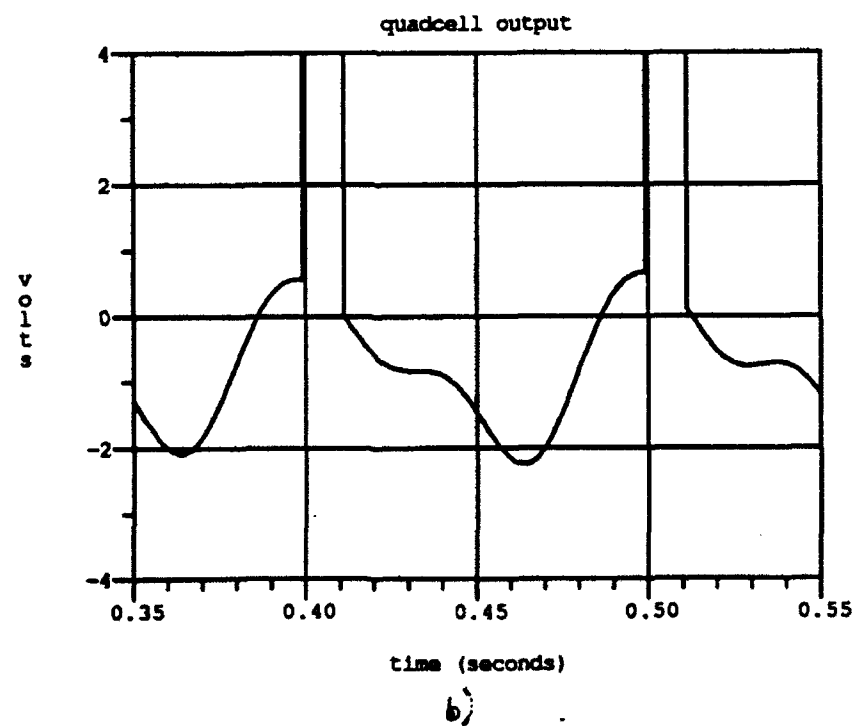
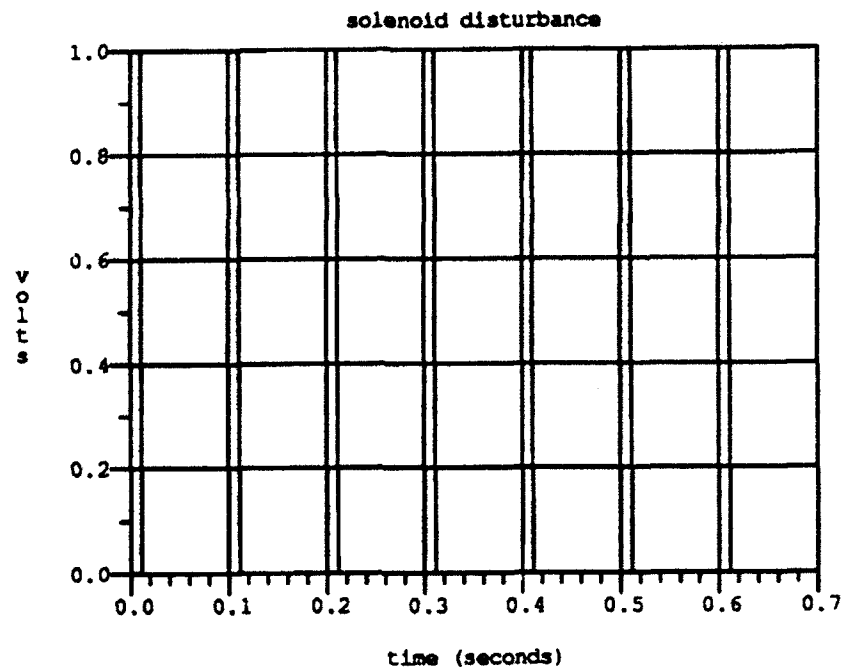


Figure 4. Time histories using mu-synthesis design.

6 Summary

A preliminary μ -synthesis control law design and some associated analyses have been carried out for the ATB-1000 test fixture. Two other SISO controllers were designed for comparison, but only the description of the designs appear here. The μ -synthesis design, when simulated with a nonlinear model of the test fixture, exhibited unacceptable limit-cycle behavior. In future work, the limits to achievable performance will be established by quantifying the key tradeoffs in terms of μ and plots like Figure 3. It is expected that the performance will be improved but still limited, for example, by a certain nonlinearity or a particular pulse frequency. The μ -synthesis methodology is well-suited for sorting out such issues.

References

- [1] G.J. Balas, "Robust Control of Flexible Structures: Theory and Experiments," PhD Dissertation, California Institute of Technology, 1990.
- [2] R.H. Cannon and D.E. Rosenthal, "Experiments in Control of Flexible Structures with Noncolocated Sensors and Actuators," *Journal of Guidance Control and Dynamics*, Vol. 7, No. 5, p. 546, Sept.-Oct., 1984.
- [3] D. Enns, "Model Reduction for Control System Design", PhD Dissertation, Dept. of Aeronautics and Astronautics, Stanford University, 1984.
- [4] C.S. Greene and G. Stein, "Inherent Damping, Solvability Conditions, and Solutions for Structural Vibration Control," Proc. of 1979 IEEE Conference on Decision and Control, December 1979.
- [5] D.E. Rosenthal, "Experiments in Control of Flexible Structure with Uncertain Parameters," PhD Dissertation, Dept. of Aeronautics and Astronautics, Stanford University, 1984.

AN INTRODUCTION TO THE TRAJECTORY PATTERN METHOD AND ITS CONTROL APPLICATIONS

F. Tangerman

Institute for Mathematical Sciences and
Department of Applied Mathematics,

J. Rastegar

Department of Mechanical Engineering,

SUNY Stony Brook, New York 11794

ABSTRACT

In this paper we introduce a new, model based, nonlinear control method, referred to as the Trajectory Pattern Method. The Trajectory Pattern Method is based exclusively on the desired behavior of the system. We show that this method is robust, allows for long term planning, and can be used to adaptively control systems with parameter uncertainty.

1. INTRODUCTION

The Trajectory Pattern Method is a new inverse dynamics based controller. The object of this paper is to explain the Trajectory Pattern Method and to show some of its potential applications in control of nonlinear systems. Such inverse dynamics based controllers have been extensively applied to the tracking control of robot manipulators and other similar nonlinear dynamics systems [1-3,13-15]. One major drawback of the presently available model based controllers is their high sensitivity to the model parameter inaccuracies and variations. In order to overcome this problem, a number of learning and adaptive control schemes [16,17] and parameter identification and calibration algorithms have been developed to compensate the effects of model parameter uncertainties [18-21]. Here, no attempt is made to present a comprehensive review of the literature.

The authors have been developing the Trajectory Pattern Method [1-3] over the past three years. To date, this method has been applied to the study of the inherent characteristics of the nonlinear dynamics of manipulators [4-5]; the problem of manipulator type synthesis for minimal high frequency vibrational excitation [6]; the study of the effects of the payload on the vibrational excitation during motion [7]; trajectory synthesis for robot manipulators for minimal vibrational excitation due to the payload [12]; the trajectory synthesis for minimal residual vibration of the tip motion of a high speed positioning machine with structural flexibility [8], a flexible beam [9], a system with joint and structural flexibility [10]; to a system with joint flexibility for minimum attainable time and minimal residual vibration [11]; and the tracking control of robot manipulators [1].

In this paper, we develop the basic structure of the Trajectory Pattern Method. The emphasis is on systems described by finite dimensional ordinary differential equations. The robustness in the presence of full and partial control is discussed. The approach can accommodate holonomic and nonholonomic constraints. In section 3, we apply this methodology to

develop nonlinear adaptive and efficient model parameter uncertainty compensation schemes. A rigorous theoretical treatment of these topics will be presented in a forthcoming paper [22].

2. TRAJECTORY PATTERN METHOD

In this section we present the Trajectory Pattern Method in some detail. We show that it is a comprehensive method for motion planning and control with open and closed loop control.

We restrict ourselves to those systems whose control is determined by ordinary differential equations. For the sake of simplicity, we only consider mostly second order systems. We first consider the case of full control.

Full Control

Consider a second order, fully coupled, differential control system on a configuration space C which is an open subset of R^n :

$$A(x, \dot{x}) \ddot{x} + B(x, \dot{x}) = u$$

u is in the input-space R^n . The state space S for this system is $C \times R^n$. We assume that A and B are smoothly varying and that for each $(x, \dot{x}) \in S$, $A(x, \dot{x})$ is an invertible $n \times n$ matrix.

Examples of such differential control systems are rigid robotic systems with actuators at all the joints. Here u is the vector of torques delivered at all the joints and the differential equation represents the balance of dynamic forces and applied forces. If the torque is delivered through a *DC* engine and is controlled by a voltage to the engine, then the voltage determines the rate of change of the torque and by differentiating the balance equation we obtain an equation of degree 3, in terms of the voltage control. The Trajectory Pattern Method extends to higher degree equations in a self-evident manner.

The Trajectory Pattern Method is a generalization of the the computed torque method, which we briefly recall. Given a trajectory $\gamma(t)$ in the configuration space C . Define the control $u(t)$ as:

$$u(t) = A(\gamma(t), \dot{\gamma}(t)) \ddot{\gamma}(t) + B(\gamma(t), \dot{\gamma}(t))$$

Then if

$$(x(t_0), \dot{x}(t_0)) = (\gamma(t_0), \dot{\gamma}(t_0))$$

and $x(t)$ is defined as the solution of the open loop control:

$$A(x, \dot{x}) \ddot{x} + B(x, \dot{x}) = u(t)$$

then

$$x(t) = \gamma(t)$$

The main idea of the Trajectory Pattern Method is that if one has a family of trajectories which accounts for all the different initial conditions, then, by applying computed torque, the system will follow for any initial condition one trajectory in the family. Usually, these trajectories are chosen so that the end condition represents a desired state of the system.

Definition: A trajectory pattern $\phi(a, t)$ over a time interval $[t_0, t_1)$ with trajectory parameters a in the space P is a smooth map:

$$\phi : P \times [t_0, t_1) \rightarrow C$$

so that for all $t \in [t_0, t_1)$, the map:

$$a \rightarrow \Phi(a, t) = (\phi(a, t), \partial_t \phi(a, t))$$

is a diffeomorphism between P and S .

Remark: For an n th order system one would consider the map

$$a \rightarrow \Phi(a, t) = (\phi(a, t), \partial_t \phi(a, t), \dots, \partial_t^{n-1} \phi(a, t))$$

Examples:

1. $C = R^n$, $S = P = R^n \times R^n$,

$$\phi((a_1, a_2), t) = a_1 t^2 + a_2 t^3$$

on the time interval $[-1, 0)$. Then

$$\Phi(a_1, a_2, t) = \begin{pmatrix} t^2 & t^3 \\ 2t & 3t^2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

All trajectories pass through 0 at time 0, with zero velocity. We sketch a few of these patterns for dimension $n = 1$ in Fig. 1.

2. $C = R^n$, $S = P = R^n \times R^n$,

$$\phi((a_1, a_2), t) = a_1 \sin(\pi t) + a_2 \sin(2\pi t)$$

on the time interval $[-1/2, 0)$

$$\Phi(a_1, a_2, t) = \begin{pmatrix} \sin(\pi t) & \sin(2\pi t) \\ \pi \cos(\pi t) & 2\pi \cos(2\pi t) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

All trajectories are periodic and pass through 0 at integer times with varying velocity, see Fig. 2.

In general, for motions that must satisfy a certain number of end conditions, an appropriate trajectory pattern is readily selected. Such a trajectory pattern can be additionally selected on the basis of some optimality criterion.

Given a trajectory pattern ϕ . It defines a control $u_\phi(x, \dot{x}, t)$ in the following manner:

$$u_\phi(x, \dot{x}, t) = A(x, \dot{x}) \partial_t^2 \phi(a(t), t) + B(x, \dot{x})$$

where $a(t) = \Phi^{-1}(x(t), \dot{x}(t), t)$.

The open loop control system defined by the trajectory pattern ϕ is the following:

$$A(x, \dot{x}) \ddot{x} + B(x, \dot{x}) = u_\phi(x, \dot{x}, t)$$

Proposition: *Given any initial condition*

$$(x(t_0), \dot{x}(t_0)) \in S$$

The solution $x(t)$ of the control-system defined by ϕ is the trajectory $\phi(a(t_0), t)$, where the parameter $a(t_0)$ equals:

$$a(t_0) = \Phi^{-1}(x(t_0), \dots, x^{(k-1)}(t_0), t_0)$$

Proof: The solution $x(t)$ satisfies the differential equation:

$$\ddot{x} = \partial_t^2 \phi(a(t), t)$$

The trajectory $\phi(a(t_0), t)$ satisfies this differential equation and has the same initial conditions at time t_0 . Therefore by uniqueness of solutions to differential equations, $x(t) = \phi(a(t_0), t)$. Q.E.D.

Remark: Observe that another way of stating this proposition is that the parameter $a(t)$ are constant and solely determined by the initial condition.

Example: Consider the simplest control equation:

$$A \ddot{x} = u$$

Consider trajectories:

$$\phi(a_1, a_2, t) = a_1 t + a_2 t^2$$

Then the map $\Phi(a_1, a_2, t)$ equals:

$$\begin{pmatrix} a_1 t + a_2 t^2 \\ a_1 + 2 a_2 t \end{pmatrix} = \begin{pmatrix} t & t^2 \\ 1 & 2t \end{pmatrix} \cdot \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$$

Therefore:

$$\begin{pmatrix} a_1(t) \\ a_2(t) \end{pmatrix} = \Phi^{-1}(x, \dot{x}, t) = \begin{pmatrix} \frac{2}{t^2} x - \frac{\dot{x}}{t} \\ \frac{1}{t^2} x + \frac{1}{t} \dot{x} \end{pmatrix}$$

and

$$u_\phi(x, \dot{x}, t) = 2 A a_2(t) = 2 A \left(\frac{-x}{t^2} + \frac{\dot{x}}{t} \right)$$

The open loop control equation is then:

$$A \ddot{x} = 2 A \left(\frac{-x}{t^2} + \frac{\dot{x}}{t} \right)$$

i.e.

$$\ddot{x} = 2 \left(\frac{-x}{t^2} + \frac{\dot{x}}{t} \right)$$

its solutions are linear combination of the functions t and t^2 .

Robustness

We are considering systems that are accurately modeled by the equations of motion. In that case, these equations of motion still depend on system parameters such as physical parameters (masses, lengths, etc.). The previous discussion then represents the ideal case: the system parameters, i.e., A and B are known exactly. The first question then is how to define the controls if the parameters are known only up to some accuracy and to determine the robustness of this method. In the simplest implementation of the Trajectory Pattern Method we attempt to control the real system (A, B) , using as control the Trajectory Pattern Method control for an approximate system (\hat{A}, \hat{B}) . This means that given the trajectory pattern ϕ , define the control $u_\phi(x, \dot{x}, t)$ as:

$$u_\phi(x, \dot{x}, t) = \hat{A}(x, \dot{x}) \partial_t^2 \phi(a(t), t) + \hat{B}(x, \dot{x})$$

Then one obtains as open loop control for the real system:

$$A(x, \dot{x}) \ddot{x} + B(x, \dot{x}) = \hat{A}(x, \dot{x}) \partial_t^2 \phi(a(t), t) + \hat{B}(x, \dot{x})$$

We now explain why this method is robust. We know the solution of this system exactly when $A = \hat{A}$ and $B = \hat{B}$, namely the trajectory pattern trajectories. By continuous dependence on parameters of solutions to differential equations, we conclude that for any compact time interval and any compact set of initial conditions, the solutions to this system are close to the trajectory patterns provided (A, B) and (\hat{A}, \hat{B}) are sufficiently nearby on a neighborhood of the region traversed by the patterns starting in this compact set of initial conditions during the time interval.

The previous robustness results are not so useful when the system runs for a fairly long time, since errors tend to build up (long term memory). Such effects need particular consideration if one wishes to regulate a system. We now discuss particular trajectory patterns which tremendously reduce the build up of long term errors and which are moreover useful when careful aiming is necessary. Those patterns focus on particular end conditions.

We say that a trajectory pattern $\phi(., t)$ on the time interval $[t_0, t_1]$ focuses at time t_1 if and only if the map $\Phi(a, t_1)$ is well-defined and has lower dimensional image.

We present examples of such focusing patterns in dimension n :

1. $\phi(a_1, a_2, t) = x_0 + a_1 t + a_2 t^2$. This pattern focuses so that at time $t = 0$ all the trajectories pass through x_0 , with variable velocity.
2. $\phi(a_1, a_2, t) = x_0 + v_0 t + a_1 t^2 + a_2 t^3$. This pattern focuses at time $t = 0$ to the point (x_0, v_0) , i.e. all trajectories pass at time $t = 0$ through the point x_0 with velocity v_0 .
3. $\phi(a_1, a_2, t) = x_0 + a_1 \sin(t) + a_2 \sin(2t)$. This pattern is periodic and focuses at time $t = n\pi$ to the point x_0 .

It is an interesting fact that the property of having focusing trajectories is robust for the Trajectory Pattern Method:

Given a control system (\hat{A}, \hat{B}) and a trajectory pattern which at time $t = t_0$ focuses to a region. Given a compact set of initial conditions then if (A, B) is sufficiently close to (\hat{A}, \hat{B}) all solutions of the control-equations with initial conditions in the compact set focus at time t_0 to the same region.

In particular, if the trajectory patterns focus on point, the trajectories of (A, B) focus on the same point. Moreover, if the trajectory patterns focus periodically, so will the trajectories of (A, B)

The mathematical explanation of this phenomenon resides in a careful study of the solutions to the singular differential equation. We will illustrate this on a nonlinear scalar example with constant leading coefficient. Consider control systems of the form:

$$A\ddot{x} + B(x, \dot{x}) = u$$

Assume that (A, B) have nominal values (\hat{A}, \hat{B}) and assume that we use as trajectory pattern:

$$\phi(a_1, a_2, t) = a_1 t^2 + a_2 t^3$$

i.e., the goal is to be at time $t = 0$ at the origin with 0 velocity. Then

$$u_\phi(x, \dot{x}, t) = \hat{A}\left(\frac{4\dot{x}}{t} - \frac{6x}{t^2}\right) + \hat{B}(x, \dot{x})$$

The open loop control system given by the Trajectory Pattern method is then:

$$A\ddot{x} = \hat{A}\left(\frac{4\dot{x}}{t} - \frac{6x}{t^2}\right) + \hat{B}(x, \dot{x}) - B(x, \dot{x})$$

This differential equation is singular at $t = 0$ and the leading order is determined by the linear singular differential equation:

$$A\ddot{x} = \hat{A}\left(\frac{4\dot{x}}{t} - \frac{6x}{t^2}\right)$$

Its indicial equation for solutions of the form t^p is the quadratic equation:

$$Ap(p-1) = \hat{A}(4p-6)$$

If $A = \hat{A}$, the solution are $p_1 = 2$ and $p_2 = 3$ (as should be the case). Consequently, when A/\hat{A} is close to one, the roots p_1 and p_2 are close to 2 and 3. But then for the original equation all solutions $(x(t), \dot{x}(t))$ which are for t near zero close to the point $(0, 0)$ actually pass through $(0, 0)$ and $x(t)$ goes to zero at least as fast as:

$$t^{p_1} \text{ or } t^{p_1} \ln(t)$$

Remark: If $p_1 < 2$ then $\ddot{x}(t)$ may become unbounded as $t \rightarrow 0$ and forces become unbounded. This can be fixed by choosing the trajectory patterns so that also the acceleration vanishes at $t = 0$. For instance:

$$\phi(a_1, a_2, t) = a_1 t^3 + a_2 t^4$$

Then the roots will be close to 3 and 4.

Constraints

We next illustrate the Trajectory Pattern Method in the presence of constraints. There are basically two kinds of constraints. Holonomic and nonholonomic constraints which give relations between the variables in the state-space and lessen the dimension of the state-space. We will discuss these briefly at the end of this section. The constraints we consider first are those which reduces the dimension of the input-space. Such constraints can for example occur in vibrational systems with finite dimensional models due to the absence of actuation in some degrees of freedom. Here the segments are divided in small segments which are considered to be rigid. At most of the joints between these segments one cannot exert any actuating torque. The joints then fall into the two classes: ones where one can exert torque and the ones where one cannot. The dimension of the input-space is then less than the dimension of the configuration space.

We assume that the equations are again second order, but that the control vector u is of lower dimension than the configuration space:

$$A(x, \dot{x})\ddot{x} + B(x, \dot{x}) = \begin{pmatrix} u \\ 0 \end{pmatrix}$$

Here u is in the m -dimensional input space. Let us assume that the configuration space $C = R^n$. We then write any vector $x \in C$ in terms of its m -dimensional component x_1 and its $n - m$ -dimensional component x_2 :

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

We then have the constraint equation:

$$A_{2,1}(x, \dot{x})\ddot{x}_1 + A_{2,2}(x, \dot{x})\ddot{x}_2 + B_2(x, \dot{x}) = 0$$

Contrary to the previous case the actual construction of trajectories is a problem. If we now assume that the matrix $A_{2,2}$ is invertible, then this constraint equation can be considered as a differential equation for x_2 in terms of $(x_1(t), \dot{x}_1(t))$, which can then in principle be solved for $x_2(t)$. Therefore, given a trajectory $x_1(t)$, and initial conditions $(x_2(t_0), \dot{x}_2(t_0))$, then the required computed torque for the system to follow the trajectory $x_1(t)$ is:

$$u(t) = A_{1,1}(x(t), \dot{x}(t))\ddot{x}_1(t) + A_{1,2}(x(t), \dot{x}(t))\ddot{x}_2(t) + B_1(x(t), \dot{x}(t))$$

For the implementation of the Trajectory Pattern Method, two related issues need to be addressed. The first is that one may also wish to control the motion of $x_2(t)$. The second is that it is in general impossible to exactly, let alone algebraically, determine $x_2(t)$, given $x_1(t)$. However, within many classes of systems, approximate explicit solutions can be obtained using algebraic techniques. There is always a trade off between the length of the time interval over which the solution is approximated and the degree of accuracy of the approximation.

In most cases of interest, the constraint equation is predominantly linear in x_2 , because x_2 , the "uncontrolled" variable should be kept small. If that is really impossible, most engineering designs would in fact add more actuators. If the constraint equation is linear,

the Trajectory Pattern Method then proceeds by algebraically solving it. If for instance $x_1(t)$ is given as a Fourier-series with a fixed period, determine the Fourier-series expansion of x_1 using the same period. The effectiveness of such a procedure for practical purposes depends on to what extent this algebraic procedure can be numerically implemented. Since the structure of the model equations is known, in many instances this can be done in real time [10]. Granted this procedure, the motion of x_2 can then be controlled by choosing trajectory patterns for x_1 .

Of course, if the constraint differential equation has constant coefficients, explicit formulae for x_2 in terms of x_1 are easily obtained.

In the case of holonomic and nonholonomic constraints, again one has to synthesize trajectories that satisfy these constraints. The method of choice is direct elimination of extraneous variables if possible. Otherwise, one can make local approximations using techniques such as implicit differentiation.

3. ADAPTIVE PARAMETER BASED UNCERTAINTY SCHEMES

Consider a class of second order differential control systems with full control:

$$A(x, \dot{x}, p) \ddot{x} + B(x, \dot{x}, p) \dot{x} = u$$

Here we assume that the parameter p ranges over a finite dimensional space of system parameters: masses, lengths, moments of inertia etc. In particular, the exact value of the parameters is usually not known and the object is to devise adaptive control schemes that compensate for this uncertainty. There are two types of approaches to this. The first type consists of devising a learning scheme which continuously makes best guesses for the values of the parameter. Such schemes are usually based on the observation that if we take any time-interval $[t_0, t_1]$ on which we assume that the values of the parameters do not change and give any input $u(t)$, then the values of the parameters can be determined/estimated by observing the resulting motion $x(t)$. If we assume that the differential equation depends on the parameters in an algebraic manner this leads to algebraic equations which can be solved to give parameter values which are consistent with the observed motion.

This approach can be implemented in the Trajectory Pattern Method as follows. Assume a nominal value \hat{p} for the parameter and consider a given trajectory pattern $\phi(a, t)$. Two distinct situations are usually encountered: regulation and tracking. By regulation we mean that eventually we want an end goal to be achieved. We assume that the trajectory patterns have these end goals and the object then is to make the trajectory parameters constant. By tracking we mean that a specific trajectory pattern described by trajectory parameter $a(0)$ needs to be followed.

First consider the case of regulation. Apply Trajectory Pattern Control with

$$\hat{A}(x, \dot{x}) = A(x, \dot{x}, \hat{p})$$

$$\hat{B}(x, \dot{x}) = B(x, \dot{x}, \hat{p})$$

If we now monitor the motion $x(t)$ and compute the trajectory parameter $a(t)$ then we have to consider two cases. The first one is that $a(t)$ is constant. In that case the system follows

the predesigned trajectory and even if $p \neq \hat{p}$ there is no reason for action. The second case is that $a(t)$ does vary. Then certainly $p \neq \hat{p}$ and one has to determine values for p which are consistent with the observed variation in $a(t)$. If $a(t)$ depends algebraically on $(x(t), \dot{x}(t))$ then one again obtains algebraic equations for the parameter p . If these equations can be solved quickly, the nominal value \hat{p} for p can be replaced by a more accurate value. Giving the corresponding trajectory control then results in the actual trajectories to be followed. This scheme is usually too optimistic because of measurement errors as well as the complexity of the algebraic equations. Therefore this scheme needs to be combined with an updating method for \hat{p} , for instance, based on Newton's method. The main point here is that $a(t) = a(t, p, \hat{p})$ and we want that $a(t)$ to tend to a constant, i.e.

$$\partial_t a(t, p, \hat{p}) = 0, \quad \partial_t^2 a(t, p, \hat{p}) = 0 \dots$$

Now explicit equations can be obtained for:

$$\partial_t a(t, p, \hat{p}), \quad \partial_t^2 a(t, p, \hat{p}) \dots$$

in terms of p and \hat{p} . If we divide the time interval in small enough intervals $[t_{i-1}, t_i]$ and choose a sequence of updates $\hat{p}(i)$ constant in the i -th interval, we can determine an equation for $\hat{p}(i+1)$ from:

$$\partial_t a(t_i) = \partial_p \partial_t a(t_i, \hat{p}(i), \hat{p}(i))(\hat{p}(i+1) - \hat{p}(i))$$

Here the left-hand side is explicitly measured, while the partial derivatives with respect to p have to be computed analytically. More equations can be obtained by taking higher time derivatives. For the case of tracking, these equations are supplemented by the requirement that $a(t)$ tend to $a(0)$.

The major advantage of the Trajectory Pattern Method in this application is that the structure of these equations is determined in advance. Therefore the equations are available in exact form and do not introduce numerical errors during computation, and is available for automated control in preprocessed form.

The other approach is based on the following idea, which we illustrate by its first order scheme. Again assume a nominal value \hat{p} for p and assume that \hat{p} is close to p : $p = \hat{p} + \Delta p$. Choose a trajectory pattern ϕ and apply as control:

$$u = u_\phi + \Delta u$$

Here Δu still needs to be determined. The basic idea is that if we choose Δu to be zero in an initial time interval, we can learn the error Δp , by comparing the actual path x_a and the desired trajectory pattern x_d . We now design Δu on the remainder of the time-interval by planning that the error between the actual path and the part tends to zero in a prescribed manner and use computed torque.

Remark: In the previous approach we chose to update p to $\hat{p} + \Delta p$ and apply the trajectory pattern control. In both approaches, a practical way to determine Δp is from Δa .

We will illustrate this technique in a nonlinear example with constant coefficients:

$$A \ddot{x} + B \dot{x}^2 + C x = u$$

Given the initial conditions, the trajectory pattern ϕ selects our desired trajectory $x_d(t)$. Assume that the nominal values for the parameters (A, B, C) are $(\hat{A}, \hat{B}, \hat{C})$:

$$A = \hat{A} + \Delta A, B = \hat{B} + \Delta B, C = \hat{C} + \Delta C$$

Denote at any time the actual trajectory by $x_a(t)$ and the error by $e(t)$:

$$e(t) = x_a(t) - x_d(t)$$

if we write u as $u_\phi + \Delta u$ then we obtain the linearized equation:

$$\Delta A \ddot{x}_d + \Delta B \dot{x}_d^2 + \Delta C x_d + \hat{A} \ddot{e} + 2\hat{B} \dot{x}_d \dot{e} + \hat{C} e = \Delta u$$

In the beginning of the motion we set $\Delta u = 0$. By observing $x_a(t)$, we compute $e(t)$ and compute $(\Delta A, \Delta B, \Delta C)$. Now choose for the remainder of the time interval a trajectory pattern for $e(t)$. Then this equation determines the adaptive feed-back $\Delta u(t)$.

The system in this example:

$$A \ddot{x} + B \dot{x}^2 + C x = u$$

has some additional structure. If x is given by a Fourier-series of size N with frequency ω :

$$x(t) = \sum_{-N}^{+N} x_n \exp(in\omega t)$$

then $u(t)$ is given as a Fourier-series of size $2N$.

Consequently, if $x_d(t)$ is given as a Fourier-series of with N and we plan $e(t)$ to be also a Fourier-series of width N then also $\Delta u(t)$ is given as a Fourier-series of width $2N$.

4. DISCUSSION

In this paper we have demonstrated the basics of the Trajectory Pattern Method. It is a feed-forward method and is based on a accurate knowledge of the system and on the designers choice of desired trajectories. We have shown that robust long time control and accurate aiming is possible using focusing trajectories. We also showed that the Trajectory Pattern Method extends so as to incorporate parameter uncertainty compensation schemes.

For practical implementations of this method one has to consider the effects of small delays in the control. Such delays occur because measurements have to be made and computations have to be performed. We mention briefly to what extent the method is robust also for small time-delays. From the theory of smooth differential equations with delay, it is known that as the delay tends to zero, solutions converge to the differential equation with delay zero. Therefore, the Trajectory Pattern Method is also robust for small enough time delays except in the neighborhood of focusing points, where the differential equation becomes singular and very near by the focusing point, the control after delay may well point in the opposite direction of what is desired. Near such focusing points the system behaves linearly and the control can be switched to simple linear controls such as PID control.

A successful implementation of the Trajectory Pattern Method generally has a strong computational component. This explains why this method was only recently developed. A careful selection of the trajectory patterns tremendously reduces the on-line computational component [1, 3, 8, 10, 12].

REFERENCES:

1. Fardanesh, B., and Rastegar J., 1992, "A new Model Based Tracking Controller for Robot Manipulators Using the Trajectory Pattern Inverse Dynamics", *IEEE Trans. Robotics and Automation*, 8, (2), pp 279-285.
2. Rastegar, J., and Fardanesh, B., 1990, "Trajectory Pattern Specific Inverse Dynamics Formulation of Robot Manipulators and its Applications", *ASME Mechanisms Conference*, Chicago.
3. Rastegar, J., and Fardanesh, B., 1991, "Inverse Dynamics Models of Robot Manipulators Using Trajectory Patterns - With Application to Learning Controllers", *Eighth World Congress on the Theory of Machines and Mechanisms*, Czechoslovakia.
4. Tu, Q., and Rastegar, J., 1991, "On the Inherent Characteristics of the Dynamics of Robot Manipulators", *13th Biennial ASME Conference on Mechanical Vibration and Noise*, Miami.
5. Rastegar, J., and Tu, Q., 1992, "On the Effects of the Operating Speed on the Dynamic Behavior of Manipulators with Rigid Links", *1992 ASME Mechanisms Conference*, Chicago.
6. Rastegar, J., and Tu, Q., 1992, "Effects of the Manipulator Type on the Vibrational Excitation During Motion", *1992 ASME Mechanisms Conference*, Chicago.
7. Tu, Q., and Rastegar, J., 1991, "A Study of the Effects of Payload on the Vibrational Excitation of Robot Manipulators During Motion", *2nd National Applied Mechanisms and Robotics Conference*, Ohio.
8. Tu, Q., Rastegar, J., and Singh, R. J., 1992, Trajectory Synthesis and Inverse Dynamics Model Formulation and Control of Tip Motion of a High performance Flexible Positioning System", *1992 Japan-U.S.A. Symposium on Flexible Automation*, California.
9. Tu, Q., and Rastegar, J., 1991, "Inverse Dynamics formulation and Control of Flexible Positioning Systems for Oscillation Free Motion", *2nd National Applied Mechanisms and Robotics Conference*, Ohio.
10. Rastegar, J., Tu, Q., Fardanesh, B., Coleman, N., and Mattice, M., 1992, "Experimental Implementation of Trajectory Pattern Inverse Dynamics Model Based Controller for a Flexible Structure", *1992 American Control Conference*, Chicago.

11. Tu, Q., and Rastegar, J., 1992, "Trajectory Synthesis for Minimum Attainable Time for Point to Point Motions of a System with a Flexible Joint", *1992 CSME Forum*, Montreal, Canada.
12. Tu, Q., and Rastegar, J., 1992, "Manipulator Trajectory Synthesis for Minimal Vibrational Excitation due to Payload", *1992 CSME Forum*, Montreal, Canada.
13. An, C.H., Atkeson, C.G., Griffiths, J.D. and Hollerbach, J.M., 1989, "Experimental Evaluation of Feedforward and Computed Torque Control", *IEEE Trans. on Robotics and Automation* 5, (9), pp. 368-372.
14. Asada, H., Ma, Z.D. and Park, J.H., 1989, "Inverse Dynamics of Flexible Robots: Feasible Solutions and Arm Design Guidelines", *ASME Robotics Research, DCS-Vol. 14*, pp. 279-288.
15. Goldenberg, A.A. and Rakhsha, F., 1986, "Feedforward Control of a Single-Link Flexible Robot", *Mechanisms and Machine-Theory*, 21 (4), pp. 325-335.
16. Miller, W.T., Glanz, F.H. and Kraft, L.G., 1987, "Application of a general Learning Algorithm to the Control of Robotic Manipulators", *The Intern. Journal of Robotics Research* 6 (2), pp. 84-98.
17. Arimoto S., Kawamura, S. and Miyazaki, F., 1984, "Bettering Operation of Robots by Learning", *Journal of Robotic Systems* 1, pp. 123-140.
18. Atkeson, C.G., An, C.H., and Hollerbach, J.M., 1986, "Estimation of the Inertia Parameters of Manipulator Loads and Links", *The Intern. Journal of Robotics Research* 5, pp. 101-109.
19. Ha, I., Ko, M. and Kwon, S.K., 1989, "An Efficient Algorithm for the Model Parameters of Robotic Manipulators", *IEEE Trans. on Robotics and Automation* 5 (9), pp. 386-394.
20. Khosla, P.K., 1989, "Categorization of Parameters in the Dynamic Robot Model", *IEEE Trans. on Robotics and Automation* 5 (9), pp. 261-268.
21. Whitney, D.E., Lozinski, C.A. and Rourke, J.M., 1986, "Industrial Robot Calibration Method and Results", *ASME Journal of Dyn. Sys., Meas., Cont.* 108, pp. 1-8.
22. Tangerman, F., and Rastegar, J., "On the Stability of the Trajectory Pattern Method", in preparation.
23. Tangerman, F., and Rastegar, J., "A Nonlinear Adaptive Uncertainty Compensation Scheme Based on Trajectory Patterns," in preparation.

LIST OF FIGURES

Figure 1: The trajectory pattern $\phi((a_1, a_2), t) = a_1 t^2 + a_2 t^3$

Figure 2: The trajectory pattern $\phi((a_1, a_2), t) = a_1 \sin(\pi t) + a_2 \sin(2 \pi t)$

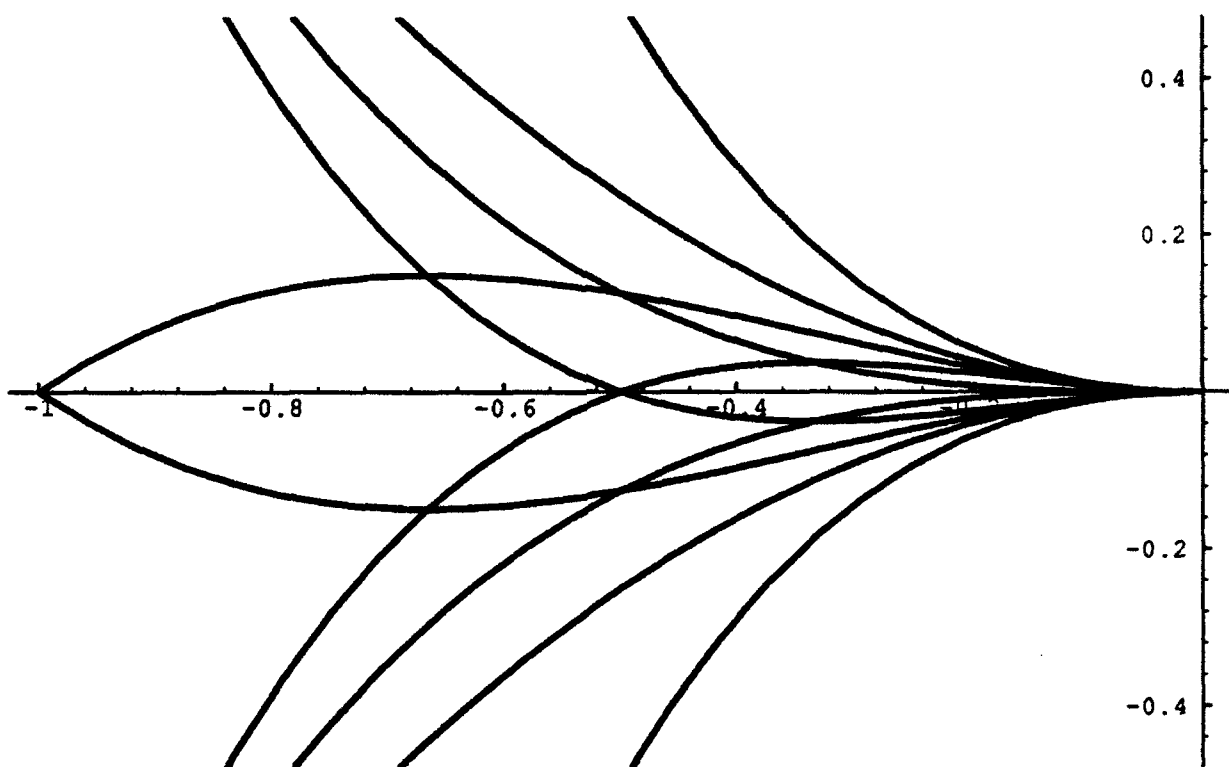


Fig 1

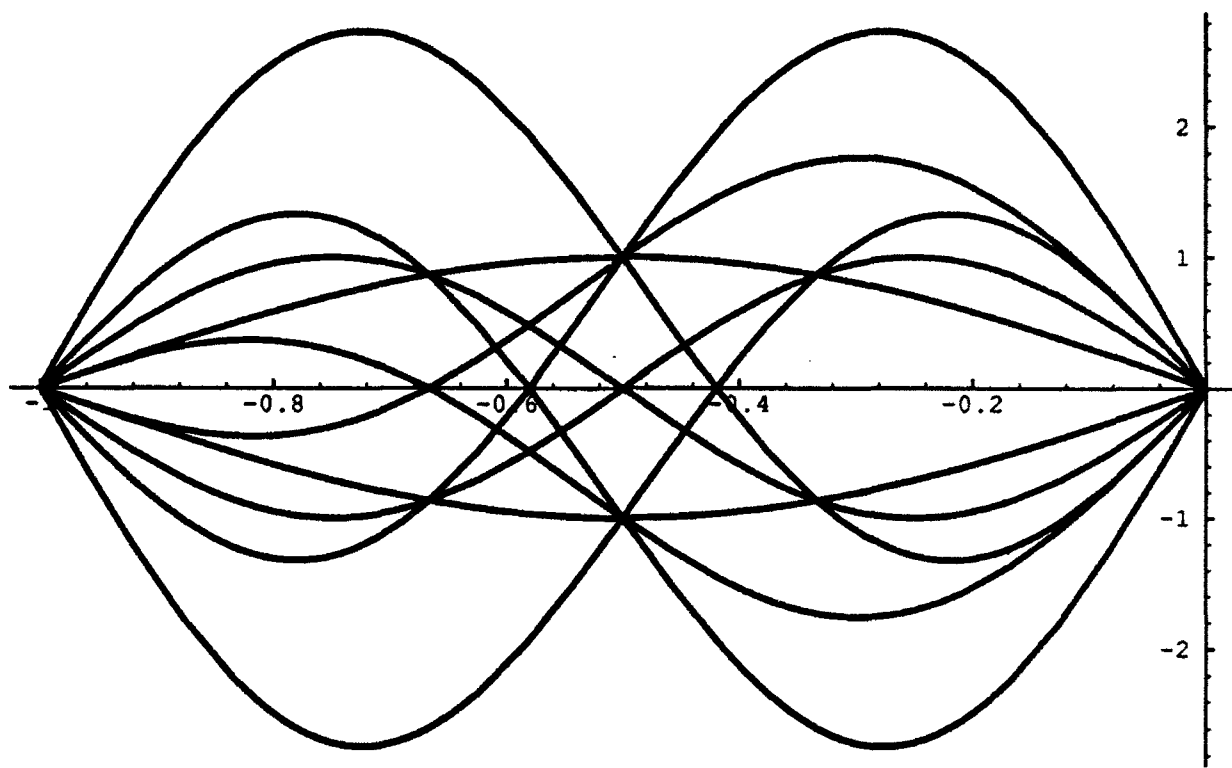


Fig 2

Visualization of Dynamic Soil-Structure Interaction Analysis

Ms. Julie Baca (Information Technology Laboratory)
Dr. Robert L. Hall (Structures Laboratory)
Mr. Donald H. Nelson (Structures Laboratory)
USAE Waterways Experiment Station, Vicksburg, MS 39180

Introduction

Advances in computer technology have enabled scientists and engineers to analyze increasingly sophisticated problems. Supercomputer access has made three-dimensional (3D) finite element modeling a tenable approach to problems once simplified to one or two dimensions for computational purposes. Such refinements in computational methods have compelled the technical community to perform prototype testing of more complex problems for the purposes of validating these computational methods. A major obstacle to overcome in this process is the inability absorb large amounts of output from analyses of complex 3D problems. Visualization techniques, which are able to graphically display analytical output in a concise manner, have become absolutely necessary for effectively handling these analyses. The U.S. Army Corps of Engineers Waterways Experiment Station (WES) has been actively employing state-of-the-art visualization techniques in support of its research mission. This paper discusses a few areas in which visualization techniques have proven beneficial in projects sponsored by the U.S. Department of Defense, Defense Nuclear Agency (DNA) and also discusses details of visualization hardware and software utilized for these analyses.

Background

DNA supports research concerning the responses of structures to dynamic loads caused by nuclear detonations. This research typically involves high pressure blast loading, soil-structure interaction, nonlinear deformations and failure mechanisms such as buckling. Results of DNA-sponsored research have led to an improvement in the ability to analyze these types of problems, however, certain structural response phenomena have not been satisfactorily analyzed by even the most complex finite element codes. There will continue to be a need for research to improve computational methods in many areas of complex structural response.

WES has been active in analytical projects for DNA which address their interests in structural response phenomena. This research has included a high precision test program so that the

response of test structures could be used to assess the accuracy of the finite element calculations. Visualization methods have proven very beneficial and in some cases indispensable for 3D finite element analysis of these problems. Visualization methods have become, and will become even more so, an integral part of the analytical process.

Visualization Techniques

Visualization techniques are graphical means of displaying large amounts of data in a manner that can be easily and accurately interpreted. Techniques such as rendering and animation lend themselves very well for displaying complex 3D finite element models and dynamic analyses involving the interaction of moving parts. Many innovative visualization methods have been advanced in recent years. Coupled with a graphics workstation, visualization methods can dramatically improve post-processing activities of computer analyses using finite element methods. In the following sections several benefits of using up-to-date visualization techniques are discussed. Unfortunately, because visualization methods are primarily in color, visual examples of these methods could not be included in this paper.

Increased Efficiency

The use of advanced color graphics to display data can substantially reduce the time required to understand analytical results allowing the researcher to more thoroughly address the problem at hand. This is especially true for analytical methods which produce considerable output. In a 3D analysis of a complex dynamic soil-structure interaction problem on a supercomputer, finite element models typically contain less than 60,000 elements to keep runtimes reasonable, i.e. less than 10 cpu hours. Computer runs such as this can produce data dump files which total several gigabytes. Using older post-processing methods a researcher would have to ignore much of this output and limit post-processing to a few chosen areas and time states. Older methods, for example black and white strain contour lines overlaid on a view of the finite element mesh, must often be examined meticulously just to see trends in an analysis. Advanced visualization methods, using color fringing to represent the same data, allow the researcher to see trends without getting lost in details. Because it is easier to spot trends, better understanding of the performance of the model as a whole is attained.

With improved efficiency, it is not necessary to impose constraints that limit post-processing to a few specified model locations and points in time simply because it is too cumbersome to interpret results. It takes only one glance to see the color coded location of peak response of the model and a second glance at the color bar to know the value. For complex geometry and responses, older methods of data processing can require several minutes of

examination just to find the locations of peak response. It is obvious that as analytical methods advance to address more complex problems, the improved efficiency attainable with advanced visualization techniques will be a necessity.

Improved Modeling

A better understanding of the performance of a model enhances efforts to improve the model. A benefit of advanced visualization is the ability to recognize modeling errors or the effect of modeling techniques used to simplify the problem. For example, often analysts purposely increase element size at model locations away from points of interest in order to reduce the problem size and decrease run time. In the finite element code DYNA3D (Reference 1) this change to coarser meshing can be achieved using a tied surface between the regions of fine meshing and coarser meshing without using transition elements. Before using advanced visualization the effect of this technique on the model performance was never fully appreciated. In problems involving shock waves passing through soil the shock front was significantly smeared and distorted at the location of the mesh discontinuity and in regions of coarser meshing. The degree of smearing and distortion was easily visible in color fringe plots of soil stress. Using older visualization methods the effect of this meshing technique did not stand out in the results. This experience highlighted the importance of ensuring that such mesh discontinuities are located well away from points of interest.

Improved Visualization

As the name implies, advanced visualization methods help an analyst to visualize the subject of an analysis and the phenomena involved. This is especially helpful in a dynamic analysis which includes large deformations and structural interaction between different parts. Animation is a great tool for checking model performance and observing how the model simulates the physics of a problem. Without animation one must mentally reconstruct the sequence of events leading up to the final results. Looking at individual frames from a dynamic analysis provide much information for doing this but putting these frames into motion yields much more. Animation allows one to visualize the time dependent aspect of data in a dynamic analysis. Phenomena such as buckling and vibration can be assessed intuitively because animation plays back the data in a natural format. There have been several occasions in which animation answered questions that had been left unanswered by examining analytical results in a piecemeal fashion.

Improved Communication

Visualization methods facilitate better transfer of information in the technical community. One of the most difficult jobs for the analyst is to share information with someone who is

not familiar with the details of an analysis. Advanced visualization can present a large amount of information in a manner that is concise and not visually overwhelming. It therefore, requires much less effort to understand a concept or a calculational result.

Visualization methods at WES

At WES, structural engineers from the Structural Mechanics Division (SMD) are assisted in developing methods to visualize analytical results by an interdisciplinary visualization team of computer scientists and engineers within the Information Technology Laboratory (ITL). In the following sections an example of how visualization methods were developed and/or modified for a specific DNA project will be described in detail. The project involved a DYNA3D finite element analysis of a dynamic soil-structure interaction and included large deformations and buckling of a buried structure. The nature of the problem required a detailed structure and a large island of soil. The resulting finite element model was relatively large, approximately 50,000 nodes, and required from 2 to 4 hours of CPU time to run on a Cray Y-MP at WES.

The initial goals of the visualization were loosely defined. Clearly, the project demanded some type of finite element compatible program capable of handling time-dependent data. In addition, the engineers desired a visual product that could serve as both an effective workroom tool as well as a means of obtaining presentation quality graphics.

The visualization team explored several alternatives in an effort to satisfy this confluence of goals. The first option explored was attempting to run DYNA3D on a graphics workstation. However, the extent of the DYNA3D data both in the number of files needed (two files per state dump with 50 state dumps) and the size of the files (approximately 3.6 megabytes each) made this approach untenable. As limitations with other alternatives continued to surface, it became apparent that a distributed computing environment was essential to gain the computational power, disk capacity, and graphics capabilities necessary for the task. Multi-Purpose Graphics System (MPGS), a software package produced by Cray Research Inc., emerged as the candidate addressing the largest number of the project's requirements. Residing on the supercomputer and a local graphics workstation, MPGS provided the distributed computing environment deemed requisite for the project. Further, because it makes no assumptions about the geometry, the software can be used with a variety of computer programs, including those for finite element analysis.

Hardware Configuration

The hardware required to visualize the finite element analysis

consisted of two distributed components, a Silicon Graphics workstation and a Cray Y-MP supercomputer, both located at WES in IITL. The workstation and the supercomputer shared the computational workload via MPGS to produce the 3D graphical images for animation. Two 25-MHz MIPS processors enabled the workstation to perform the local graphics manipulations: the workstation display has a resolution of 1280 by 1024 pixels with 24 bit-planes of color. The workstation and the supercomputer communicated over a standard Ethernet network connection using TCP/IP; MPGS uses sockets to move data between the workstation and the supercomputer. These network connections permitted compute-intensive tasks to be performed on the supercomputer and the results downloaded to the workstation from within MPGS.

Software Issues

Data Format

MPGS can process up to three categories of data for visualization. First, the required geometry data describe nodes, lines, elements, and solids. The geometry data, which may comprise several files, represent only the initial geometry of the structure. The optional scalar data contain multiple scalar data for each node. The optional vector data contain one 3D vector value for each node. For time-dependent scalar or vector data, multiple files are necessary, each representing a state dump of output data.

The DYNA3D binary data files requiring translation to MPGS format, each approximately 3.6 megabytes in size, contain first a time stamp and control information, then the geometry data. The time-dependent data follow, including nodal coordinates, and scalar values. Multiple time-dependent data sets cannot be contained in one file. For each new state dump, DYNA3D opens a new file and writes the nodal and scalar data for that state dump.

Data Translation

Cray Research Inc. furnishes software that translates DYNA3D data files to suitable MPGS format. Specifically, three translators generate the geometry, scalar, and vector data files, respectively. The Cray Research translators, however, addressed only limited cases and required two principal modifications (Figure 1). First, each translator was modified to read and convert DYNA3D data residing in multiple files. Second, each of the scalar and vector translators processed only brick elements, neglecting shell and beam elements. Both translators were modified to also process shell elements. Enhancements to support beam elements are currently being pursued.

VISUALIZING MISTY PORT - HARDWARE AND SOFTWARE COMPONENTS

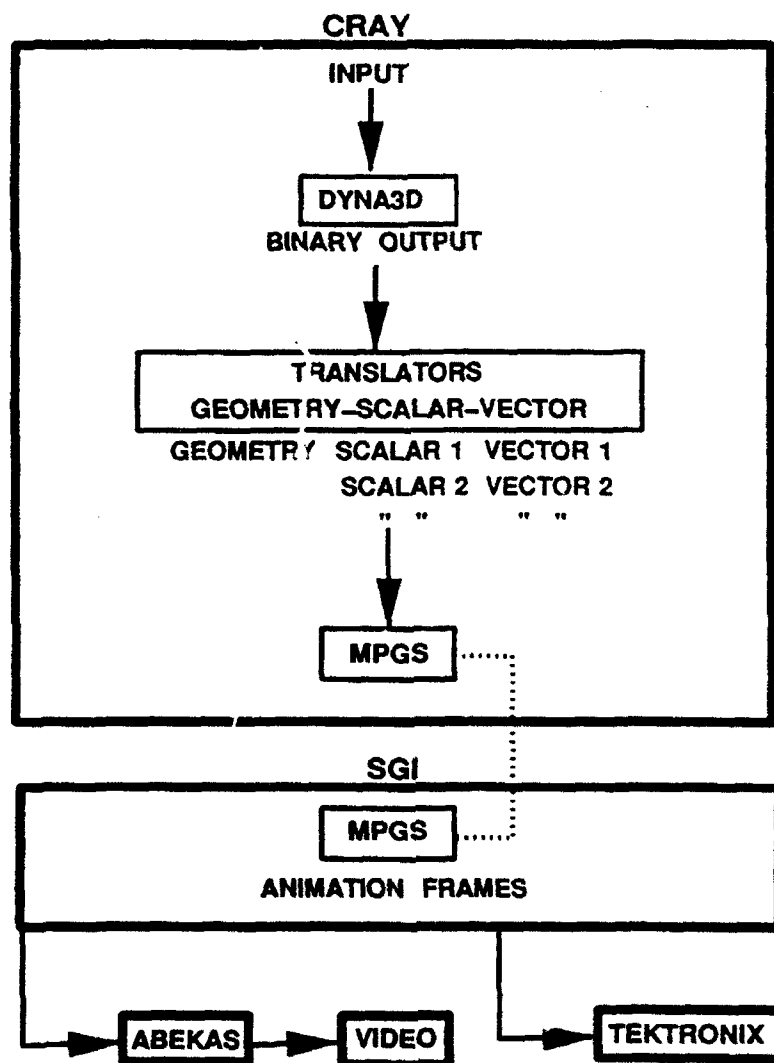


Figure 1 Hardware and Software Components

Viewing Model Results

Once the initial geometry was obtained, it became necessary to isolate specific parts of interest to the animation and analysis. As detailed, MPGS allows multiple element types with attributes tied to each part. Exploiting this feature to "turn off" all elements except those representing the structure permitted deformations and stresses to be viewed clearly. To view deformations, time-dependent displacement computations were performed on the supercomputer, converted to graphical elements, and downloaded to the workstation via the built-in MPGS communications features.

To further examine geologic influences on the structure, vertical, horizontal, and shear stress contours were computed and rendered. The rendering of the stresses illustrated problems with tied surfaces connecting zones of fine discretization with coarse discretization in the soil adjacent to the structure. The information led to rediscrctizing the soil-structure system and moving the tied surfaces away from the structure. This resolved problems with the buckling of the structure due to approximations associated with the tied surfaces. In other analyses, visualization has similarly proven useful not only for interpreting results but also for highlighting problem areas with the finite element model.

Conclusions

Through the cooperative endeavor, the structural engineers realized many of their goals, some initially unarticulated. Examination of deformed shapes and stresses from the analysis in a dynamic presentation provided details of soil-structure interaction which had never been visualized before. Visualization provided the opportunity to see the characteristics of shear friction between the soil and the structure and the role of the soil in determining the buckling mode of the structure. This demonstrates that complex structural systems can be more easily solved when visualization tools are provided to assess numerical models and analytical results.

ACKNOWLEDGMENTS

The investigations were sponsored by the U.S. Department of Defense, Defense Nuclear Agency under the direction of Major John McDugald. Investigations were conducted by the U.S. Army Engineer Waterways Experiment Station, Structural Mechanics Division. Permission from the Chief of Engineers to publish this paper is gratefully acknowledged.

The contents of this paper are not to be used for advertising, publication, or promotional purposes. Citation of trade names does

not constitute an official endorsement or approval of the use of such commercial products.

REFERENCES

1. Hallquist, J.O., and Benson, D.J., DYNA3D User's Manual, Revision 3, Lawrence Livermore National Laboratory, July 1987.
2. Nelson, D.H., and Hall, R.L., "Dynamic Nonlinear Finite Element Modeling of Buried Thin-Walled Cylinders", 62nd Shock and Vibration Proceedings, 1991.
3. Grimsrud, Anders. MPGS, A Distributed Multi Purpose Graphics System, Version 3.5, Cray Research Inc., October 1990.
4. Peterson, J.W., Bogart, R.G., and Thomas, S.W., University of Utah, Department of Computer Science, Salt Lake City, UT, Utah Raster Toolkit.

ADVANCED COMPUTER MODELING OF METEOROLOGICAL
EFFECTS UPON ARTILLERY PROJECTILE FLIGHT

Abel J. Blanco
U.S. Army Atmospheric Sciences Laboratory
White Sands Missile Range, NM 88002-5501

Sherrill J.H. Edwards
electroMAGIC Tools
Organ, NM 88052-0490

ABSTRACT. The trajectory of an artillery projectile can rarely be characterized by a single set of meteorological measurements (MET Message); however, current practice is to use a single MET message to derive the final aiming adjustments necessary to hit the target. Examining the meteorological effects upon projectiles dynamically through modeling allows a better understanding of the probability of hitting the target the first time. A criterion is developed for utilizing multiple MET messages and applying a solo MET message at particular projectile trajectory locations to obtain a high probability of success. The Advanced Battlefield Environment Artillery Model (ABEAM), a U.S. Army Atmospheric Sciences Laboratory version of the U.S. Army Ballistic Research Laboratory General Trajectory (GTRAJ) model, is the advanced model environment for developing this examination. ABEAM is used in prototyping new artillery meteorological techniques for current and future artillery systems. Preliminary results reveal significant aiming adjustment differences computed from a proposed MET correction and the current MET doctrine. Since multiple MET messages allow a more realistic description of the battlefield atmosphere, the proposed methodology can significantly improve the accuracy of artillery predicted fire.

INTRODUCTION. For each mission the field artillery has to quickly respond with accurate, massed, and surprise fire. A highly efficient fire for engagement at extended ranges demands a modern artillery fire direction. The U.S. Army Atmospheric Sciences Laboratory (ASL) is using applied mathematics and innovative computing algorithms to demonstrate the possible artillery predicted fire accuracy improvement. This note presents a brief description of a modern meteorological (MET) adjustment technique; artillery accuracy comparisons; MET scenario results; and simulated paired statistics of significant accuracy improvements.

The current MET aiming adjustment for predicted fire applied to aiming future extended target ranges is not sufficient most of the time. Since the atmospheric conditions are not homogeneous along the projectile trajectory, the current method of adjusting artillery fire with MET data collected from a dedicated station is no longer valid. With these extended ranges the expected MET bias increases and continues to be a major contributor in the total artillery error budget (see Lillard et al, 1990). The lengthening flight time has significantly increased the unit MET corrections. To continue accurate artillery fire capability at the new ranges an improved MET final aiming adjustment is required. ASL is researching new procedures for allowing a more realistic battlefield atmosphere into the artillery fire direction center. A module that allows multiple MET message inputs and solo message application at

particular trajectory portions was incorporated into the U.S. Army Ballistic Research Laboratory GTRAJ model. This modification provides an advance model environment for testing the worth of proposed MET adjustment techniques. The new configuration, the ABEAM was validated and tested by Blanco and Edwards (1991).

The Project PASS (Prototype Artillery Subsystem) field data (see Blanco and Traylor, 1976) is utilized for the testing of proposed MET adjustment procedures. Because of time constraints only 2 of the 20 data days are used to present results from the proposed MET adjustment as compared to the current MET correction. These 2 selected data days contain the most variable weather conditions. A preliminary conclusion presents the expected artillery accuracy improvement afforded by proposed multiple MET messages and selection of a solo MET message along the launch, apogee, and target projectile flight location.

MODERN ARTILLERY MET ADJUSTMENT. In aiming artillery the fire direction center converts weapon and target location into firing data and utilizes a ballistic simulation model that requires MET data input for the final firing angles solution. A single computer MET message (METCM) provides the atmospheric description. The header in figure 1 identifies the MET station location and the body lists values for the following MET parameters—wind direction, windspeed, virtual temperature, and atmospheric pressure. The raw MET data are averaged into horizontal, vertically layered computer zones listed on the first column. The last computer zone is line number 26 with the bottom and top at 19 and 20 km above the surface. The current fire direction application assumes that the computer MET message data is horizontally homogeneous along all portions of the projectile trajectory. This is not a valid assumption, especially at extended ranges.

COMPUTER MET MESSAGE					
For use of this form, see FM 6-15; the proponent agency is TRADOC					
IDENTIF. : OCTANT	LOCATION	DATE : TIME : DURATION	STATION : MDP		
CATION :	LaLaLa LeLeLe	: : (HOURS)	HEIGHT : PRESSURE		
	or		(10° M) : MB°		
METCM : Q	zzz zzz	YY : GGGGG : (G)	hhh : PpPpPp		
METCM : 1	347 987	15 : 140 : 0	842 : 079		

ZONE VALUES					
ZONE HEIGHTS (METERS)	LINE NUMBER	WIND DIRECTION (10° M)	WIND SPEED (KNOTS)	TEMPERATURE (1/10°K)	PRESSURE (MILLIBARS)
	23	ddd	PPP	TTT	PPP
SURFACE	00	004	007	2710	0970
200	01	010	030	2711	0967
500	02	000	023	2713	0937
1000	03	003	032	2688	0891
1500	04	008	033	2640	0836
2000	05	039	037	2620	0783

Figure 1. Artillery computer MET message.

Note that there are specific MET effects along different portions of the projectile trajectory. For example, a rocket assisted round or a thrusting rocket require a tailored launch cross wind adjustment. This effect is propagated throughout the whole flight time, but the interaction takes place during initial motion and burnout time - the launch portion. The time from burnout to the warhead event defines the coasting portion. And the time from

the warhead event and impact defines the target area. Depending on the relative MET station location and the wind flow, the dedicated MET message may not be representative of the launch, coasting, or target area MET. For this reason the artillery needs a modern MET message that can provide more realistic data along these trajectory portions.

Using the standard munitions aerodynamic coefficient tables extended to Mach 5 (see Lieske, 1990) to represent future artillery capability, the GTRAJ model was utilized to compute four projectile trajectories. Figure 2 details the projectile position at a 2-sec flight time interval. The apogee or coasting area for the 15-km trajectory is defined as that portion above the 2 km where the projectile spent over 20 sec flight time. Approximately one-half of all extended range trajectory flight time is spent in the apogee portion. For the coasting rocket or shell the most important MET observations are required between the preapogee and postapogee portion. Normally, windspeed, the major MET contributor, is at a maximum for the trajectory during apogee; thus, the MET effect has its best chance to deviate the projectile away from the initial aim point at this apogee portion. And for the accurate delivery of submunitions the fire direction must use realistic target area MET. Parachute-delivered submunitions, chemical warheads, and wind gliding warheads require accurate wind measurements in the target area.

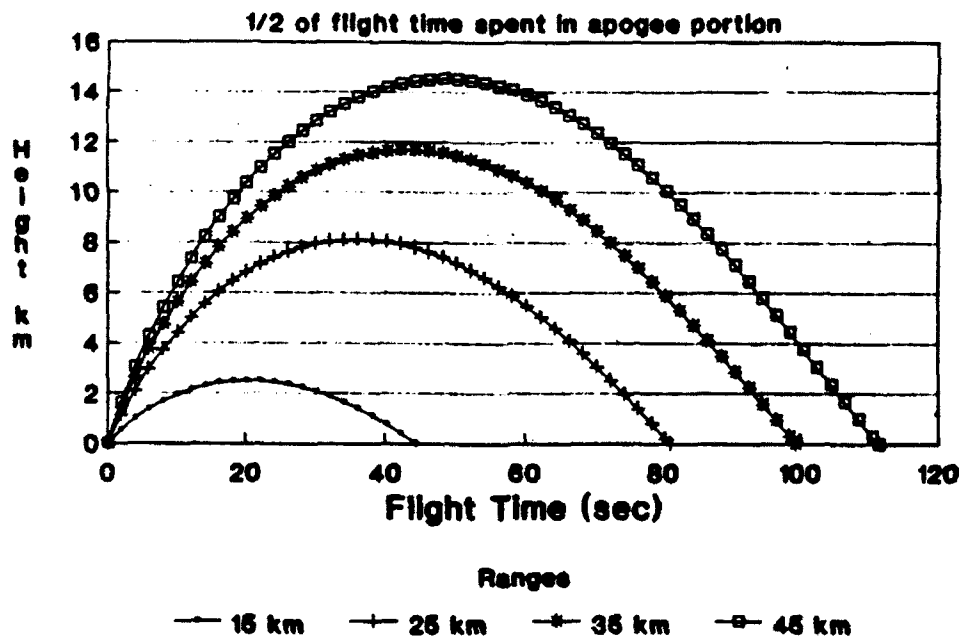


Figure 2. Long range artillery trajectories.

Centralizing all available battlefield METCM and tailoring a METCM for a particular trajectory portion provides a more realistic atmospheric description. The fire direction center can receive multiple computer MET messages and select the most representative for application at the different portions of the trajectory. This procedure may be automated in the fire direction center or manually performed by its operator. The same North Atlantic Treaty Organization (NATO) standardized format is used and the only change is to report all available METCMs to the fire direction center and

follow criteria in selecting an appropriate solo METCM for the corresponding trajectory portion.

ARTILLERY ACCURACY COMPARISONS. The worth of the proposed MET aiming techniques can be evaluated through statistical analysis of simulated and actual impacts. Figure 3 identifies a self-propelled howitzer and a target location normalized impact. The solid line ellipse represents the accuracy and MET day-to-day dispersion afforded by simulated no MET corrections. This procedure of using the standard MET conditions to represent current conditions is known as the "cold stick" method. In this case the bias errors (largest displayed) represent the total uncorrected MET effect. Method A represents the actual surveyed MET bias errors afforded by the current doctrine. If the dedicated METCM was representative of what the shell experienced, then the dashed line ellipse would be centered about the target with only the MET day-to-day and hardware precision dispersion. Method B represents the simulated improvement afforded by using MET data that is more representative of the MET experienced by the projectile as it traverses through the different portions of its flight trajectory. The smallest bias errors are not zeroed because of the always present time staleness and space displacement between the MET measurement and projectile location. However, the MET day-to-day precision error is significantly improved because the proposed technique consistently reduces the MET bias.

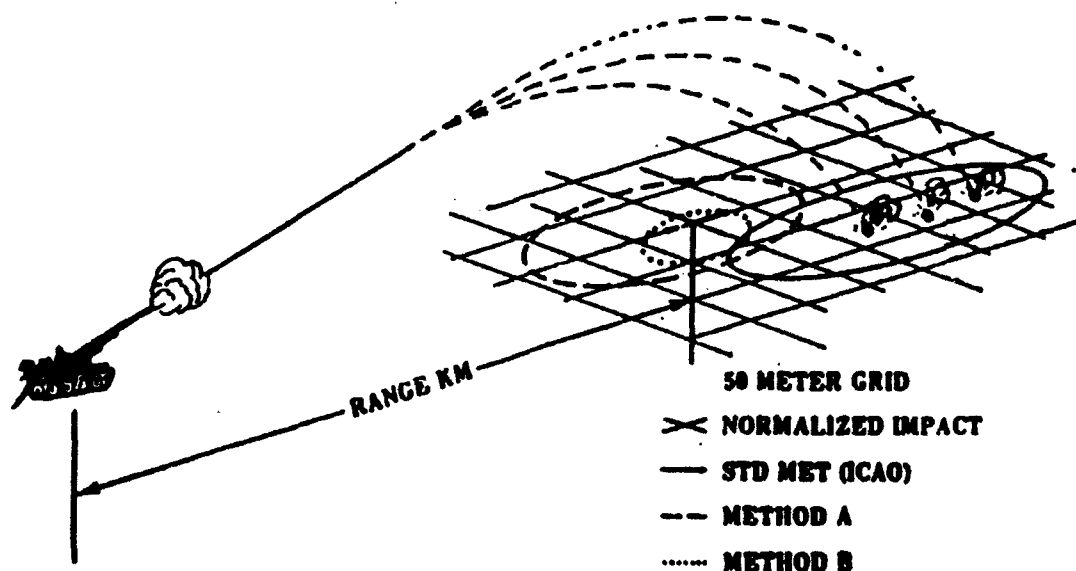


Figure 3. Expected MET effects on artillery fire.

By modifying the above methodology one can extract the same comparison analysis while only using simulated results; however, there are no absolute target misses. The paired differences between methods A (doctrine) and B (proposed) are the statistics required. Through simulation one can locate the gun position and target location so that the projectile trajectory apogee traverses through an area where actual measurements are recorded. An experiment is then designed to compare simulated impacts using a single METCM

and simulated impacts using a solo METCM at the launch, apogee, and target trajectory portions.

Schematically, figure 4 represents the procedure followed in deriving these impact dispersions. For a given MET scenario the single METCM and ABEAM multiple messages are inputs to the GTRAJ model for deriving a pair of simulated impacts. Available replicates for each scenario are averaged to represent statistical results. If the paired differences are small then the single METCM is representative of the three launch, apogee, and target areas. This was not experienced as stated in the following sections.

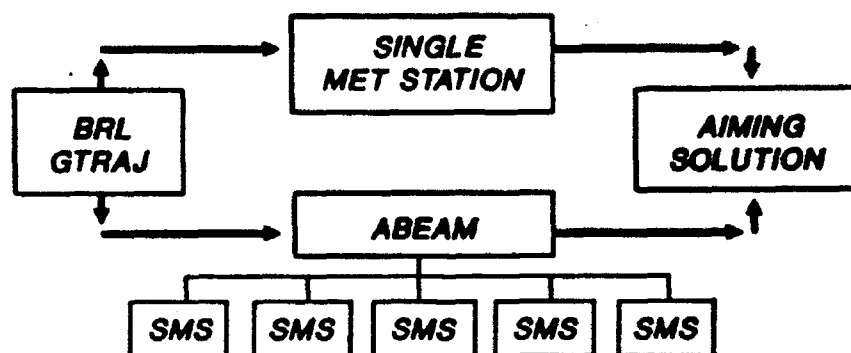


Figure 4. Schematic chart for accuracy comparisons.

MET SCENARIO RESULTS. During November and December 1974 a comprehensive ballistic MET field experiment was completed at White Sands Missile Range, New Mexico. Simultaneous rawinsonde data was collected (see D'Arcy 1977), and all the data is formatted in the NATO METCM format (see Field Manual FM 6-16, 1983). Figure 5 identifies the location of the 10 MET stations with a star. During this experiment actual firing was aimed and fired on a 14-km target range. The selected MET scenario involves only five MET stations simultaneously releasing rawinsonde balloons to achieve a multilocation model of the MET effects for an example battlefield environment.

Through simulation one can relocate the gun position and identify a target such that the simulated apogee flight will traverse over a measuring station. For example, figure 6 presents the MET scenario for the five simultaneous rawinsonde releases and a simulated 18-km range firing. At this target range the apogee is less than 3 km above the surface. The concatenated arrows represent the 1015 wind conditions at each of the six artillery computer zones included under the projectile's apogee. This top view displays the balloon drift and the gun/target line of fire. Each arrow represents the wind vector at the line number listed in figure 1.

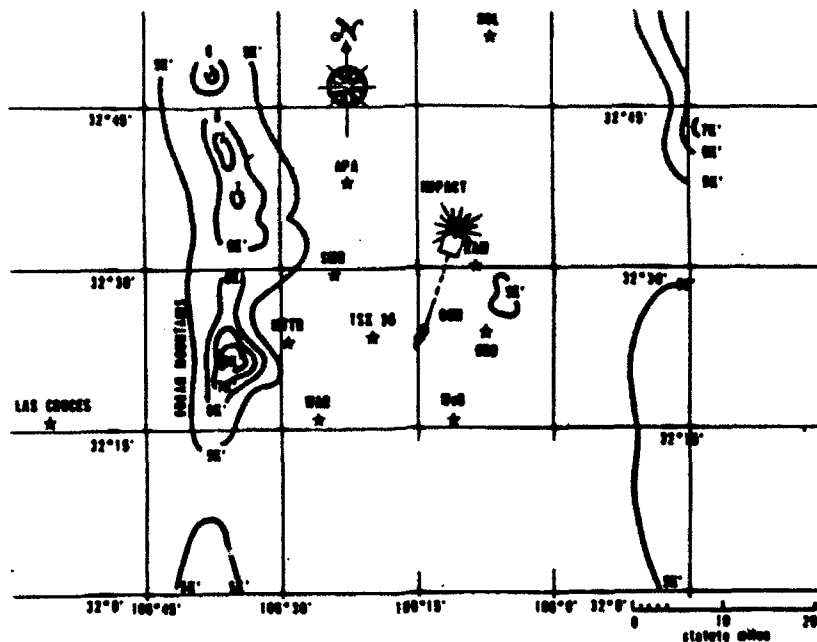


Figure 5. Field experiment at White Sands Missile Range, New Mexico.

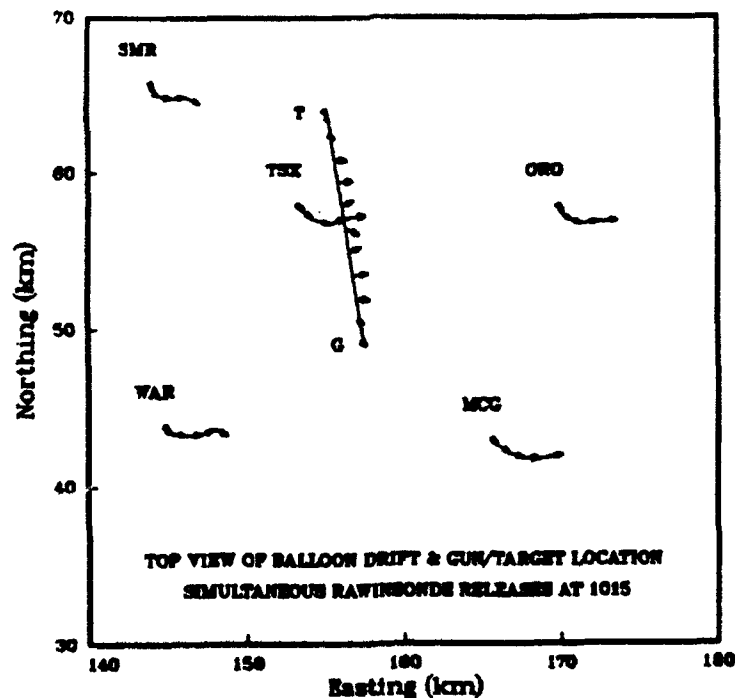


Figure 6. Space displacement MET effects.

The current doctrine is represented by superimposing the WAR (site on Martin Luther King Road, formerly War Road) METCM (dedicate station) along the flight trajectory. Data from zone 1 is applied at the launch and target areas because of the atmospheric homogeneity assumption. As stated in the above section the most important trajectory portion for a coasting shell is the apogee. Comparing the apogee (line 6) wind vector from WAR and TSX (tower

site) MET stations reveals a better understanding of the space variance effect between simultaneous wind measurements. Following the current doctrine the gun is aimed with WAR data, but the shell experienced data measured at TSX. TSX reports a total crosswind vector at apogee, while WAR reports both cross and range wind components. The largest expected miss is along the range because current doctrine required a higher quadrant elevation to compensate for the WAR reported head wind. In reality there was no head wind at apogee; therefore, the current method of aiming involves a long miss in the range, and since the WAR crosswind does not match the TSX cross wind there is a left miss in the cross component. So even with real time data because of space separation between the measurement and the point of application there will be an expected error. The lengthening of the apogee flight time will continue to increase expected errors because the atmosphere is not always homogeneous.

The time staleness between the MET measurement and the time of application presents the largest expected error. Figure 7 displays the concatenated arrows for the 1415 wind conditions except for the data collected at the dedicated WAR station. For some battlefield constraint the gun is now aimed and fired using the 1015 data or with a 4-hr old METCM. Comparing these results with those from figure 6 reveals that for this time span the wind profiles shifted from westerly to northerly direction (blowing from). Examine the wind vectors at line 6, the apogee of our simulated firing. The gun is aimed with the same WAR cross and range wind components described in the above scenario. However, in this case, the wind experienced by the shell is more in the range component and less in the cross component. Both WAR and TSX station report a head wind for the range component, but for the cross component they report significantly different winds. By simulating impacts using these METCMs, one can define the value of expected errors.

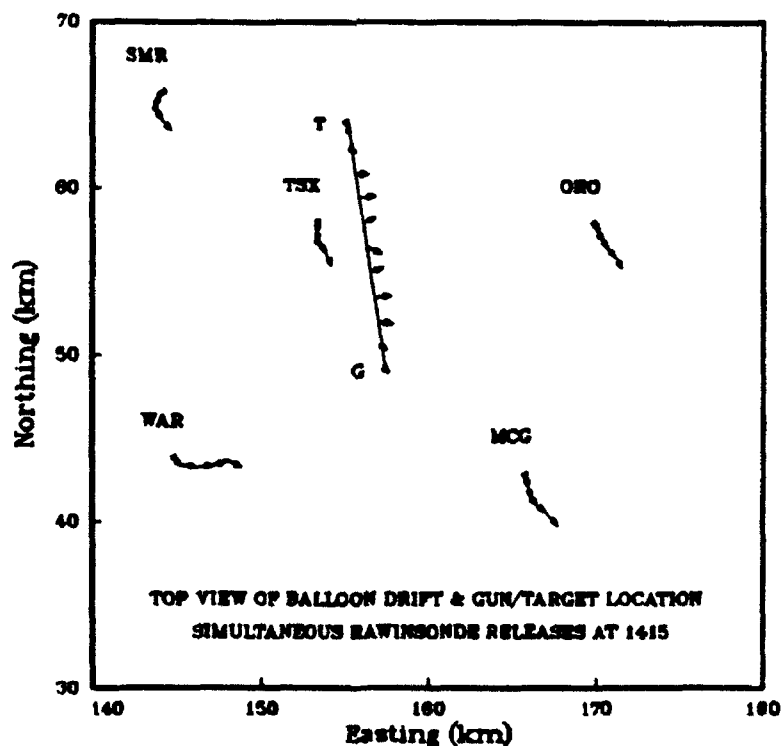


Figure 7. Time staleness met effects.

SIMULATED PAIRED STATISTICS. The artillery aiming MET effects due to space and time variabilities have been presented with no assigned accuracy errors results. Since this note presents no actual firing comparisons, only statistical results can be reported. Following the above description and only using the 2 most variable weather days of the 20 data days a set of replicates was defined. The 2 days contain five and four simultaneous releases every 2 hours. For the space variability there are nine replicates for firing from WAR over TSX towards SMR (small missile range). In representing the statistical sample for the time staleness variability, one needs to pair the results from these nine replicated into 2- and 4-hr old staleness. The total 2-hr stale paired results are 4 and 3 for each corresponding day. Similar results for the total 4-hr stale paired results are 3 and 2. Table 1 lists the paired statistics comparing simulated impacts using the current doctrine of a single dedicated MET station (WAR) and the proposed multiple METCMs applying only selected solo METCM at the critical trajectory portions (launch-WAR, Apogee-TSX, and Target-SMR).

TABLE 1. PAIRED STATISTICS BETWEEN DEDICATED AND PROPOSED METCM.

MET SPACE VARIABILITY EFFECT					
	BIAS(m)		FLIGHTS	SIGMA(m)	
	Range	Cross	n	Range	Cross
0 hrs	93	37	9	149	43
MET TIME VARIABILITY EFFECT					
2 hrs	293	102	7	204	86
4 hrs	375	156	5	312	159

With an increase in replicates, the statistical results (especially the event to event sigma results) will stabilize and be more representative of a general conclusion. However, from these preliminary results one can summarize a trend in variability and a relative accuracy effects between space displacement and time staleness. The 100-m rms space bias for an 18-km target range is too much of a difference to ignore in aiming artillery. These results represent real-time MET corrections and demonstrate that a significant improvement is very possible. The 310-m rms time bias for 2-hr staleness is also not acceptable in aiming artillery. Battlefield data must be centralized and fused into improved and modern MET correction techniques. The 406-m rms time bias for 4-hr staleness is definitely not acceptable. Current MET aiming adjustment can be significantly improved.

SUMMARY. ASL is performing applied mathematics and developing computing algorithms that allow dynamic simulated examination of meteorological effects upon projectiles traversing through a battlefield atmosphere. Since only software product improvements are required, the artillery MET aiming

adjustment can significantly be improved, thereby enhancing the first-round-hit probability. ABEAM is the advanced model environment used to examine new artillery MET correction techniques.

Preliminary results demonstrate the expected MET space and time variability effects on aiming extended range artillery. Paired statistics between the current doctrine and the proposed technique contain too much of a difference. These statistics are summarized as follows: 100 m for space variability, and 310 m (2-hr stale) and 406 m (4-hr stale) time variability. The atmosphere is not homogeneous, especially along the extended artillery firing ranges. If the U.S. Army wishes to continue with accurate and effective artillery fire at these new ranges, the current method of adjusting for nonstandard MET conditions needs to be improved. ASL is investigating several possibilities, and this note describes a model development quantifying the expected MET errors. More replicates to increase the statistical sample size will be defined using all 20 PASS data days.

In conclusion, all battlefield METCM need to be centralized. Data fusion and tailoring for launch, apogee, and target area applications need to be investigated. Analytic functional approximations, optimum interpolation, and mesoscale modeling can be used to define a solo composite METCM. More frequent battlefield MET measurements are needed. Artillery balloon-borne sensor data need to be augmented with remote sensor data.

REFERENCES

- Blanco, Abel, and Sherrill J. H. Edwards, 1991, A multiple meteorological station version of the BRL general trajectory model, Proceeding of the 1991 Battlefield Atmospheric Conference, U.S. Army Atmospheric Sciences Laboratory, White Sands Missile Range, NM, pages 395 - 404.
- Blanco, Abel, and Larry Traylor, 1976, "Artillery meteorological analysis of project PASS," ECOM-5804, U.S. Army Atmospheric Sciences Laboratory, White Sands Missile Range, NM.
- D'Arcy, Edward, 1977, "PASS 500mb rawinsonde data volume I and volume II," ECOM-DR-77-4, U.S. Army Atmospheric Sciences Laboratory, White Sands Missile Range, NM.
- Field Manual FM 6-15, August 1983, Field Artillery Meteorology, Headquarters, Department of the Army, Washington, DC.
- Lieske, Robert, 1990, "Maximum ordinates for extended Range Ordnance," Memorandum SCLBR-LF-T(340), U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD.
- Lillard, Robert, et al, 1990, Personal Communication, TRADOC, U.S. Army Field Artillery School, Fort Sill, OK.

Prepared for:
Tenth Army Conference on Applied Mathematics and Computing
U.S. Military Academy, West Point, New York, June 16-19, 1992

EXPERIMENTAL TECHNIQUES FOR SCIENTIFIC DATA INTERPRETATION

Charles S. Jones, PE *

Julia A. Baca **

* Civil Engineer
Scientific Visualization Center
USACE Waterways Experiment Station
ATTN: CEWES-IM-MS (Jones)
3909 Halls Ferry Road
Vicksburg, MS 39180-6199
sjones@wes.army.mil

** Computer Scientist
Scientific Visualization Center
USACE Waterways Experiment Station
ATTN: CEWES-IM-MS (Baca)
3909 Halls Ferry Road
Vicksburg, MS 39180-6199
jbaca@wes.army.mil

Abstract

Techniques for visualizing scientific data have advanced remarkably in recent years. These techniques remain inadequate, however, for interpreting certain classes of problems, such as those including multivariate or complex, time-dependent data. In addition to the requirements of existing problems, the scope and size of problems in the near future will extend into the teraflop range, demanding innovative solutions for interpreting results. Investigations of two experimental technologies for data interpretation, virtual reality and data sonification, are presented in this paper.

1. Introduction

The traditional scientific computing environment consisted of a large mainframe computer or supercomputer supporting a substantial user base, connected via dial-up asynchronous lines and dumb terminals. Typically, large simulation programs were run in batch mode. To some extent, this model continues to exist due to the high cost of the central resources. Enhancements to this environment, however, have significantly extended the capabilities available to modern users.

Communications are faster and more reliable. Dial-up lines have been replaced by higher bandwidth networks. Ethernet, with a bandwidth of 10 megabits per second, and Transmission Control Protocol/Internet Protocol (TCP/IP) are the current standards for interconnecting systems [9, 10]. Distributed computing and scientific visualization require even higher transmission capacities. These requirements are being met with a combination of new networking architectures, protocols, and transport media. For example, the fiber distributed data interface, FDDI, is replacing Ethernet due to its bandwidth of 100 megabits per second.

Workstations have supplanted dumb terminals on the researcher's desktop as a means of interacting with the central computer. The workstations vary from personal computers to high- performance graphics workstations. These systems possess significant computational and storage resources which rival those of first generation supercomputers. Now, client/server applications and distributed computing environments allow clusters of these systems to work together to solve large, complex problems.

Such an environment, in addition to making large numerical simulations computationally feasible, has fostered the development of more sophisticated techniques necessary for interpreting simulation results. Visualization software, integrating the computational power of supercomputers with specialized graphics hardware, has extended the set of interpretation tools available to the researcher. Two-dimensional line contours have been expanded to surfaces. Surfaces may have constant value (iso-surfaces) or may be shaded, displaying multiple values using a continuous range of colors. In addition, researchers can examine vector data using particle traces or through time-dependent animations on their workstations.

Clearly, the interpretive techniques reviewed above represent a significant advancement beyond methods available only five years ago. Nonetheless, certain classes of problems require interpretive solutions for which current techniques remain inadequate. Examples in-

clude studies with multivariate data as well as complex time-dependent problems. Research indicates that technologies engaging a fuller realm of human perception, incorporating such senses as hearing and touch as well as sight, can be effective for interpreting these classes of problems [14]. In addition to the requirements of existing problems, the scope and size of problems in the near future, extending into the teraflop range, will demand innovative solutions for interpreting results. The remainder of this paper examines two experimental technologies for data interpretation, virtual reality and data sonification, and discusses the results of our investigations.

2. Experimental Data Interpretation

2.1. Virtual Reality

Virtual reality has generated profound interest in the technical community for its potential in analyzing scientific data. Employed for such a purpose, it seeks to engage the researcher more fully in the interpretive process, allowing interaction with the data through viewing, touching, and actually moving in the physical space or "world" of the data. While the technology is receiving a great deal of attention presently, certain concepts embodied in virtual reality date back to the 1960's when Ivan Sutherland developed the first head-mounted display system [24].

The original head-mounted display consisted of a helmet containing a pair of CRTs with left-eye and right-eye views, adjusted by computer according to the user's head movements. Researchers at NASA Ames Research Center refined the head-mounted display to use liquid crystal displays (LCDs) and electronic sensors, making it practical and affordable [8].

Two companies exemplify the commercial viability of virtual reality, VPL Research Inc., and Fake Space Labs. The companies utilize similar hardware configurations but with some variation. VPL has chosen to use the head-mounted display in their EyePhone System. The display employs two color LCDs with a pixel resolution of 360 x 240. Motion is detected by the 3Space Isotrak Sensor System developed by Polhemus Navigation Sciences [27]. The 3Space system contains two components, a source and a sensor. The source generates two hemispherical electromagnetic fields, each of which has a radius of approximately 33 inches. The sensor, mounted on the headset, continuously transmits its spatial location to an A/D converter, which then transmits this information to the computer.

The Binocular Omni-Orientation Monitor (BOOM) system offers an alternative to the

VPL head-mounted system. Manufactured by Fake Space Labs, the BOOM is an extension of a prototype developed at the NASA Ames View Laboratory [4]. The BOOM supports two small CRTs, each with a pixel resolution of 720 x 486, on a counterbalanced yoke attached through six joints to a base. The apparatus uses optical encoders located in the joints to determine spatial position and orientation.

A VPL DataGlove is used in both systems as the means for directly interacting with the virtual world. The DataGlove provides two types of information. First, spatial information is generated using a Polhemus 3Space Isotrak system, described above. Second, information concerning the hand, ie. gestures, is determined through the attenuation of light transmitted along custom optical fibers which span the length of each finger on the glove. These fibers are quite sensitive, so that even such precise gestures as American Sign Language can be accurately transmitted to the computer.

2.1.1. Working with Virtual Reality: Observations

Although the VPL EyePhone System and the Fake Space Labs BOOM system share certain similarities, they differ notably in several respects, particularly the display technology and the software interface for creating virtual worlds. The ensuing discussion presents a comparative analysis of the two systems, based on our experiences. A report on the focus of current efforts as well as comments on future directions conclude the discussion.

2.1.2. Hardware Configuration

The display technologies of the two virtual reality systems present a distinct choice to the potential user. The VPL system offers the authenticity of color, while the BOOM provides increased detail due to the higher resolution display. For simulation, color is a necessary component of realism. Conversely, data interpretation benefits from the higher resolution. Eventually, high resolution color displays will be cost-effective when virtual reality becomes generally available and economies of scale can be achieved. One must also consider the advantages of the VPL head-mounted display versus the counterbalanced viewing apparatus of the BOOM. The headset imparts a feeling of being immersed in the virtual world, whereas the BOOM is more accessible for scientific data exploration.

2.1.3. Software Interface

More significant than the differences in the two display technologies are the dissimilarities in the software interface for creating virtual environments. The VPL system provides a program called Swivel to create virtual worlds [22]. Essentially a drawing program, Swivel allows the user to create objects and place them in the virtual world, with dimensions of 1000 x 1000 x 1000 units. Individual constraints and relationships between objects can be specified. Though it is possible to detect collisions between objects, interference checking is not automatic; hence, creating solid, impassable objects requires explicit testing for collision.

Posing another restriction, Swivel provides no translators for importing models, such as CADD. For these cases, the user must completely reconstruct the model, using Body Electric, a programming language specific to Swivel [3]. Externally developed codes cannot be linked into the system. Consequently, the VPL software provides a rather closed programming environment.

No software is provided with the BOOM. Working code can be obtained, nonetheless, from various organizations which have a BOOM. In general, such software is based on functions originally developed at NASA Ames [4]. These functions read information from the headset and the DataGlove and control the stereo mode of the display. This relatively low level of functionality allows total flexibility for creating virtual worlds: functions for displaying the desired data must be externally developed. This makes it possible to view essentially any type of data which can be drawn on the computer, including numerical simulation results, physical or CADD data. Thus, the BOOM provides a more open, flexible programming environment than that offered by the VPL system.

2.1.4. Applications of Virtual Reality

Our goal is to apply virtual reality to the problems associated with data interpretation. These problems differ from those for simulation in which realism, in the realm of the virtual world, is the dominant requirement. Employing virtual reality to analyze data should enable the researcher to perceive the data in new ways, discerning structure that is difficult, if not impossible, to detect using traditional computer displays.

To date, we have created three models for examination using the BOOM. The first project entailed importing a three-dimensional finite element model of steady-state flow through a channel. For the second project, a volume model was created using data col-

lected from accelerometers buried in soil. Finally, the flight of a projectile was examined. Each of the projects involved differing types of data, yet in each, a need existed for the scientist to interact with the model. For example, in studying flow through the channel, it is difficult, using two-dimensional displays, to observe the helical flow patterns that develop in the curves. Employing particle traces in virtual reality, these details can be more readily observed.

Our attitude towards the use of virtual reality for data interpretation remains ambivalent. Virtual reality provides an intuitive model for moving around in a three-dimensional world. However, visualization techniques must be adapted to facilitate exploring the data from a perspective of immersion in the virtual world, rather than the conventional approach, wherein the researcher is positioned outside that world. New paradigms must be developed, similar to the shift involved in parallel versus sequential programming.

2.2. Data Sonification

Researchers have been interested in the applications of sound in computer technology for the last decade. Interest in the area has increased as sound synthesis techniques have improved, making it possible to produce digital audio signals in real time. Much of the initial research in the area was motivated by the need to employ sound in the development of computer-human interfaces for the visually impaired [16]. Subsumed within this larger goal was the desire to use sound to interpret data, also called data sonification. However one chooses to categorize the research, the literature indicates common issues to be addressed and shared findings upon which to draw to effectively employ sound in computer-human interaction.

Among the fundamental issues which emerged from the initial investigations, it became evident that the human auditory system possessed unique attributes which could be used more fully in developing the computer-human interface [1]. First, analagous to its visual counterpart, the human auditory system can perceive certain physical dimensions of sound, such as pitch, volume, and duration. These lower level physical dimensions can be isolated and manipulated to present information similar to the way in which the visual attributes of an object such as shape, size, or color may be used [1, 5]. Secondly, at a higher level, human hearing represents a continually open channel and can thus function in the background, detecting information without requiring full conscious attention [13]. Audio cues have been used for some time to develop auditory warning systems which monitor aircraft or nuclear power plants.

Early studies also revealed that certain kinds of information are possibly better understood aurally than visually, such as time-varying, logarithmic, and multivariate data [1, 19, 26]. Finally, initial research demonstrated that any efforts to employ sound in computer-human interaction must address certain fundamental questions: What are the most appropriate mappings of information to sound and how can these mappings be determined? Also, how do human auditory and visual perception interact and how can this be maximally used to convey information [16, 18]?

2.2.1. Current Sound Technology

Lack of necessary hardware was a genuine impediment to early research efforts involving computer-generated sound. As sound synthesis techniques have improved, the applications of sound in computer technology have accelerated. This section briefly reviews current technology for generating computer-synthesized sound.

The Musical Instrument Digital Interface or MIDI, is a standard adopted by the music industry for interfacing electronic sound synthesis and processing equipment to computers. MIDI can be described as something akin to the printer language Postscript [20]. However, instead of sending page-description instructions, MIDI conveys music-description signals to a synthesizer. The standard has made a wider range of equipment available to the researcher.

It should be noted that prior to development of MIDI synthesizers, some experimentalists performed sound synthesis in software [17]. The early sound synthesis languages (Music N languages) were modular and flexible, but could not produce sound in real time [15]. MIDI synthesizers overcame this limitation, implementing the sound synthesis algorithms in hardware to produce digital audio in real time. However, because the hardware is pre-programmed, MIDI synthesizers alone do not offer the flexibility of software synthesis and require additional software and hardware components to perform experimental research.

Digital signal processors (DSPs) offer another alternative for sound synthesis. DSPs are microprocessors specifically designed for high-speed digital signal processing operations. Implementing the sound synthesis algorithms in software to be executed on one or several DSPs in parallel can yield the flexibility of software synthesis and the real-time response of MIDI synthesizers.

2.2.2. Current Data Sonification Research

Current research efforts have attempted to address the various issues raised in the initial investigations. In one such project, researchers at the National Center for Supercomputer Applications have employed a sound specification system to test the effectiveness of sonification in assisting researchers to analyze complex data [7]. The research team explored the problem of generating optimum data-to-sound mappings by testing what they term both "abstract" and "data-related" mappings. To realize the goals of the project, the investigators employed a sound specification system to analyze data sets from selected NCSA-produced scientific visualizations. The sound specification system incorporated both a general-purpose hardware component and an object-oriented software component. The project culminated in the production of sonifications which augment the selected visualizations. Results of the project indicated the significance of cognitive factors in determining effective data-to-sound mappings. The investigators point out, however, that more thorough exploration and testing of data mappings is required to develop sonification as a data interpretation technique [7].

Researchers at the University of Lowell have examined the effectiveness of sound in multivariate data interpretation, particularly the interaction between visual and aural perception in the interpretive process [14]. The project has entailed construction of an exploratory visualization tool, Exvis. The tool seeks to exploit the human ability to perceive texture, both visually and aurally. The developers argue that a texture, made up of numerous discriminable elements, permits deliberate analysis of individual elements, yet also, without requiring deliberation, provides an overall impression of the data [14]. Each data sample is represented visually by a stick-figure graphical unit called a glyph or "icon" [14]. The data parameters determine the attributes of an icon, such as the size or position of its parts. Variations in shape, size, spacing, or orientation create textural gradients or contours, which indicate to the viewer structures of potential significance in the data. Each icon has auditory as well as visual attributes. An icon may generate a single tone or other sound, depending on values assigned to its attributes. Multiple icons sounding simultaneously produce auditory textures similar to visual textures.

2.2.3. Applications of Sound Technology

Comparing the two studies, the NCSA project adopted a more general focus by employing a sound specification system to manipulate audio attributes. Combining such an approach with the Exvis emphasis on multivariate data appears to be optimal for employing sound to interpret complex data. Presently, we are in the initial stages of investigating a sound specification system and possible data mapping techniques.

Several studies are being evaluated for the application of data sonification techniques. Those involving multivariate data sets from both physical and numerical systems appear particularly well-suited. Examples include the Chesapeake Bay study and groundwater investigations: employing sound while moving through three-dimensional vector and scalar fields could yield valuable insight. A final application would be the development of user interfaces to scientific programs for the visually impaired.

3. Conclusions

Scientific visualization has extended the boundaries of traditional graphics. Nonetheless, new methods must be developed which address the interpretive requirements of more complex, time-dependent systems. This paper has examined two possible technologies, virtual reality and data sonification. Each offers potential benefits not offered by conventional methods. Through ongoing studies, we expect to determine the viability of these technologies for exploring scientific data.

4. Acknowledgements

The authors would like to acknowledge the contributions from the staff of the Scientific Visualization Center and the individual researchers who provided input for this paper. Permission was granted by the Chief of Engineers, U.S. Army Corps of Engineers, to publish this information.

The contents of this paper are not to be used for advertising, publication, or promotional purposes. Citation of trade names does not constitute an official endorsement or approval of the use of such commercial products.

5. References

- [1] S. Bly, *Presenting information in sound*, Proceedings of the CHI 1982 Conference on Human Factors in Computer Systems, pp. 371-375, 1982.
- [2] S. Bly (Ed.), *Communicating with sound*, Proceedings of CHI '85 Conference on Human Factors in Computer Systems, pp. 115-119, 1985.
- [3] Body Electric Manual v. 3.0, VPL Research Inc., 1991.
- [4] S. Bryson and C. Levitt, *The virtual windtunnel: an environment for the exploration of three-dimensional unsteady flows*, Proceedings IEEE Conference on Visualization, 1991.
- [5] W. Buxton, *The use of non-speech audio at the interface*, Tutorial #10, Proceedings of the CHI '89 Conference on Human Factors in Computer Systems, pp. 2.1-2.15, 1989.
- [6] E. Carterette and M. Friedman, Hearing, Handbook of Perception, Volume IV, New York: Academic Press, 1978.
- [7] A. Craig and C. Scaletti, *Using sound to extract meaning from complex data*, Proceedings of the SPIE Conference, E. Farrell, chair, SPIE: San Jose, 1991.
- [8] S. Ditlea, *Inside Artificial Reality*, PC Computing, November 1989.
- [9] Department of Defense, *Military Standard Internet Protocol*, MIL-STD-1777, August 12, 1983.
- [10] Department of Defense, *Military Standard Transmission Control Protocol*, MIL-STD-1778, August 12, 1983.
- [11] E. Evans and J. Wilson (Eds.), Psychophysics and Physiology of Hearing, New York: Academic Press, 1977.
- [12] J. D. Foley, *Interfaces for advanced computing*, Scientific American, pp. 127-135, 1987.
- [13] W. Gaver, *Auditory icons: using sound in computer interfaces*, Human Computer Interaction 2(2), 167-177, 1986.
- [14] G. Grinstein, R. Pickett, and M. Williams, *EXVIS: An exploratory visualization environment*, Graphics Interface '89, pp. 254-261, 1989.
- [15] G. Loy and C. Abbott, *Programming languages for computer music synthesis, performance, and composition*, Computing Surveys, Vol. 17, NO. 2, pp. 235-265, ACM, New York, 1985.
- [16] D. Lunney and R.C. Morrison, *High technology laboratory aids for visually handicapped chemistry students*, Journal of Chemical Education, 58(3):228-231, 1981.
- [17] M. V. Matthews, The Technology of Computer Music, The M.I.T. Press, Cambridge, 1969.

- [18] D. L. Mansur, *Graphs in sound: a numerical data analysis method for the blind*, MS thesis, UCRL-53548, Lawrence Livermore National Laboratory and University of California, Davis, 1984.
- [19] J.J. Mezrich, S. Frysinger, and R. Slivjanovski, *Dynamic representation of multivariate time series data*, Journal of the American Statistical Association, 79(385):34-40, 1984.
- [20] PostScript Language Reference Manual, Adobe Systems Incorporated, 1986.
- [21] J.G. Roederer, Introduction to the Physics and Psychophysics of Music, Second Edition, New York: Springer-Verlag, 1975.
- [22] RB2Swivel User Interface Documentation, VPL Research Inc., 1990.
- [23] S. Smith, D. Bergeron, and G. Grinstein, *Stereophonic and Surface Sound Generation for Exploratory Data Analysis*, Proceedings of the CHI '90 Conference on Human Factors in Computer Systems, pp. 125-132, 1990.
- [24] I. E. Sutherland, *Head-mounted three-dimensional display*, Proceedings Fall Joint Computer Conference, 1968.
- [25] E.M. Wenzel, P.K. Stone, S.S. Fisher, and S.H. Foster, *A system for three-dimensional acoustic 'visualization' in a virtual environment workstation*, Proceedings IEEE Conference on Visualization, pp. 329-337, 1990.
- [26] E. S. Yeung, *Pattern recognition by audio representation of multivariate analytical data*, Analytical Chemistry, 52(7):1120-1123, 1980.
- [27] 3Space User's Manual, Polhemus Aerospace and Electronics Company, 1987.

Flow Computation about Army Projectiles & Missiles Using Multi-Zone Grids on Parallel Computer Architectures

Dr. Nisheeth Patel*

Mr. Jerry Clarke

Mr. Monte Coleman

U.S. Army Ballistic Research Laboratory

Attn: SLCBR-SE-A

Aberdeen Proving Ground, MD 21005-5066

1. INTRODUCTION

There were three major goals of this paper. The first was to utilize the massively parallel Connection Machine, CM, for computation of flow about Army projectiles and missiles using a zonal code. Multi-zone grids on the order of several million points have been used to model real-world three-dimensional geometries with discontinuities. This allows breaking up of a complex computational domain into blocks of simple grids. Overlaid multi-zone grids were considered for the study. Since the quality of the grid has a direct impact on the quality of the solution, the grid is a dominant factor.

We have also retained the complex boundary conditions associated with practical applications. In particular, the boundary conditions were neither simplified nor hard coded inside the code for computational efficiency. They are set up in modular blocks that make it easy to use the same code for a variety of applications.

Our second goal, again emphasizing practical applications, was to compare performance of the whole application to the code kernel. The importance of the communication overhead on the CM, and communications issues related to the single-zone grid with multi-zone applications are addressed.

Our third goal was to demonstrate the usefulness of a remote visualization capability developed in conjunction with the massively parallel computations. Applications involving millions of grid-points on a remotely located CM require the transfer of enormous amounts of data to a local machine for visualization. This requires special consideration when one is interested in capturing real-time dynamic or unsteady behavior of the flow solution.

2. CONNECTION MACHINE

The CM is a single instruction multiple data (SIMD) parallel computer. A full size Connection Machine of this class has 65,536 ($=2^{16}$) bit serial processors, 16 processors to a chip. The chip also contains router circuitry for inter-processor communication. The CM has one floating-point chip for every 32 processors. A "node" consists of two processor chips, a floating point chip and memory. Thus a full size CM has 2048 nodes. The CM located at the Army High Performance Computing Research Center (AHPCRC, University of Minnesota) has 1024 nodes or 32,768 bit serial processors.

Because of the SIMD nature of the CM it is not possible to overlap computation with communication. Thus communication, if not implemented properly can significantly degrade the overall performance. The CM has a regular grid communication feature within its hypercube topology. This grid or "NEWS" communication is designed in such a way that grid neighbors are assigned to hypercube neighbors in the communication network. This allows every processor, in parallel, to pass data to its neighbor, all in the same direction. Since in our finite-difference application the communication pattern is regular, we have used the above described grid communication in the code. The cost of grid communication is typically very competitive with the cost of the basic arithmetic. On the other hand, random access or long-distance inter processor communication is far more expensive than regular grid communication. Although it is possible to pass data from each node to all four neighbors simultaneously, using a special micro-coded communication

primitive, this has not been explored because it is not considered a general purpose feature at the present time.

The CM located at the AHPCRC has 128 Kbytes of memory per processor for a total of 4 Gbytes of memory. It has a SUN front end, consisting of two 4/490 systems, to provide compiles, loads and communications. It also has a 10 Gbyte parallel data vault storage system which provides disk storage accessible at a rate of 24 MBytes/second. This consists of 42 SCSI disks operating in parallel and the special software libraries and utilities needed to use the system.

3. NAVIER-STOKES ALGORITHM

Since the architecture of the SIMD CM is well suited for an explicit type method, an explicit multi-stage finite-difference method has been adopted [Ref. 1]. The code solves Navier-Stokes equations in generalized coordinates with no thin layer assumption. The three-dimensional Navier-Stokes equations in generalized curvilinear coordinates can be found in [Ref. 1]. We can write the equations as:

$$\frac{\partial Q}{\partial t} + \frac{\partial(E - S)}{\partial \xi} + \frac{\partial(F - T)}{\partial \eta} + \frac{\partial(G - R)}{\partial \zeta} = 0$$

where Q is the vector of the conserved variables, E, F, G are the inviscid flux terms and S, T, R are the viscous terms. The code is capable of performing both steady and unsteady computations. For steady flows, a local time stepping convergence acceleration scheme is used.

We have added adaptive artificial dissipation terms to the equations. These terms are required to damp out high frequency oscillations associated with the odd-even decoupling in central-difference schemes. Also, the artificial dissipative terms are required to capture shock and contact discontinuities without undesirable oscillations. An adaptive blend of second and fourth differences has been used for artificial dissipation. The dissipation terms have been improved by a directional eigenvalue scaling [Ref. 2] which has been found effective for highly stretched grids in both inviscid and boundary layer/wake regions.

For a time accurate solution, the explicit scheme requires a limited time step size that must be determined by the numerical stability criterion of the scheme. The time step size is normally determined such that the Courant-Friedrichs-Lewy (CFL) condition is a minimum over all grid cells. However, convergence to the steady state solution can be accelerated by sacrificing the time accuracy of the scheme and advancing the solution at each cell in time by the maximum possible local time step in that cell. Local time stepping is also well suited for parallel processing because it allows concurrent computation of the time step size in all cells and avoids long-distance communication. The code allows the use of multi-zone overlaid grids. The multi-zone grid has the potential to reduce the grid generation complexities and to improve the quality of the grids associated with complex body geometries. Since the application involves grid points of the order of millions, the grid must be saved on the data-vault, a mass storage device, for fast parallel I/O in the code. If the grids and solution fields are not saved on the data-vault, the I/O time will be prohibitive for both visualization and application.

For simple body shapes, such as a body of revolution, the grid generation has been carried out on the CM itself and grid files have been saved on the data-vault. For complex body shapes requiring the use of a more involved grid generation code, grids have been generated on a local machine and the binary grid files transferred to the CM. These grids are placed on the data-vault by a pre-processing code on the CM using different file names for each grid zone.

The solution process requires the computation of flux vectors that includes artificial dissipation from the dependent variables. This can be done concurrently for all interior grid points in a given zone. In the case of a multi-zone grid, the above process is carried out for each zone until all zones are finished. Then the solution vectors are obtained by solving the first stage in parallel for all interior points of a given zone. Again the process repeats for any additional zones. Next, appropriate boundary conditions are updated, each in parallel. The major concerns when applying the complex boundary conditions, which take longer, is that processors representing interior grid points do not participate in any useful work. Thus boundary conditions must be carefully considered. The inter-zone data exchange is done next for overlaid zones. Since each grid zone is separately mapped on the CM, involving different sizes, a substantial amount of long-distance communication may be involved in

the exchange process. Although no computation is involved in this exchange process, the communication over-head is impressive. This cycle repeats for the next stage. Variants of this cycle can be explored by applying boundary conditions and/or inter-zone exchanges less frequently in each stage, to achieve better performance.

4. FLOW APPLICATION

Depending on the application, the computational domain consists of either a single-block or a multiple-block grid. Each block consists of a large number of grid points. For code implementations, the CM has been configured as a 3-D grid with one (virtual) processor per grid point and each processor connected to its nearest neighbors. The code was designed in such a way that each grid-block has its own local variables. This approach allowed us to map each grid block on the entire CM. Each grid block should contain a sufficient number of grid points (in multiples of the physical processors) to give a high virtual processor ratio for efficient computation. That is, each physical processor will process several grid points of a single block grid. Thus for a multi-zone grid, storage requirements per physical processor can become very large.

The inter-processor communications for each grid block is relatively efficient because of the previously mentioned dedicated "NEWS" network. Applications involving multi-block grids typically have grid blocks of different sizes. Thus there is substantial "long-distance" communication involved in inter-block data exchange. This is one of the major issues involved in comparison of a multi-block application with a single-block application. Hence, long-distance interprocessor communication is the first important issue. Because the CM is SIMD, it is not practical to overlap computation with the required interprocessor communication and thus the overhead generated by communication has a direct effect on performance.

During the computation of boundary conditions, un-needed processors will not be performing useful work. Since our goal is to maximize the rate of useful computations, the implementation of boundary calculations is the second important issue.

For some applications, the boundary stencil can be made part of the interior stencil by using an array that has desired factor values on the appropriate boundaries and interior factor values corresponding to the physical position in the computational field. The overhead imposed by the inclusion of additional terms when solving the interior and boundary points simultaneously is relatively small on the CM. Hence the application of difference stencils to virtual processors representing boundary points can be performed concurrently with interior points. Also the Dirichlet and periodic type of boundary conditions are easy to implement on the CM. For many applications, the type of boundary conditions required is rather complex. There is little or no similarity in the difference stencils between the interior and boundary points. This is because of the nature of the boundary conditions or higher order approximations applied to the boundary conditions. Also, the boundary conditions are evaluated after the interior points have been updated. Thus depending on the type of boundary conditions, separate evaluation of boundary conditions impose an overhead on a SIMD massively parallel architecture that mostly depends on the type of application involved and the complexity of the boundary condition. Although all boundary points in a given boundary condition can be executed in parallel, still a large ratio of virtual processors might not be doing useful work. So the type of applications and the type of boundary conditions used play an important role in comparing performance with other machines.

A number of applications have been computed to validate the accuracy of the Zonal code on the CM. Both 2-D axisymmetric and 3-D flows have been solved and compared with the experimental data and results from Cray computations. The parallel processing effort at BRL was initiated in late 1983 with the acquisition of a 64 processor Denelcor HEP MIMD machine which provided an initial platform for numerical experiments with parallel algorithms. Some of the techniques implemented in the Connection Machine code, such as multi-zone data exchange, local variables for each zone, and data management were dealt with experimentally on that computer. Also programming experience with the Cyber-200 series using explicit vector directives contributed to the CM code, because it had many directives similar to CM Fortran. With the availability of the Crays, the code has been validated and applied to real world problems. The code has consistently sustained a parallel efficiency of about 98% for 3-D applications on 4 and 8 processor Crays [Ref. 3].

We have tried to satisfy our primary aim of performance by using efficient numerical schemes and

simple data management based on modular building blocks for flexibility. The code has been written from scratch for the CM, working with one module at a time in order to simplify the development and enhance portability. Once the blocks are identified and the parallel directives are implemented, then an efficient code is developed by putting these modules together in an appropriate way. Thus we have been able to port the parallel code efficiently from one computer to another [Ref. 4]. The code has modules that are capable of computing aerodynamic forces and coefficients such as drag and pitching moments acting on the body. These coefficients along with the surface pressure distribution and velocity profiles (when available) are compared with the experimental data.

We now mention some of the application studies that have been performed using the zonal code. All of the following studies used multi-zone (block) grids involving 2 to 15 zones. The effects of projectile nose bluntness has been reported in [Ref. 1]. References [5] and [6] present detailed discussion on studies of spike-nosed HEAT (High Explosive Anti-Tank) projectiles. The non-steady behavior of the flow is interesting and real-time visualization may provide more insight of the flow behavior. The viscous flow computations have been performed to study missiles with delta wings at various angles of attack [Ref. 7]. The formations of leading edge vortices and their effect on the wing pressure distribution has been investigated. The projectile with a cylindrical wrap-around fin, which allows efficient use of the available space, has been investigated in [Ref. 8]. Computations for a fully integrated engine including inlet, combustor and nozzle for the National Aerospace Plane or X-30 are described in [Ref. 9].

In the remainder of this section we show a sampling of flow computations that has been performed on the CM at the AHPCRC using the zonal code. First we consider the Army M549 artillery shell. This is the 155mm projectile with a flat nose and a flat base. Our interest is in the computation of supersonic flow and determination of static aerodynamic coefficients. Because of the sharp edged configuration at the nose-forebody junction and at the boattail-base junction, a three zone grid has been used. The 3-zone grid for the M549 is shown in Figure 1.

The projectile shape consists of a flat nose, forebody section of ogive and cylinder and aftbody section of the boattail and flat base. The zonal grid preserves the actual corners in the nose and base area. The zonal grids are overlaid by a single cell with the neighboring grids without any mismatch and thus inter-zone boundary conditions only require data transfer at the interfaces. A non-reflecting boundary condition was imposed on the outer boundary [Ref. 1]. This approach allows setting of the outerfield close to the body. From the computational point of view, it allows the removal of a relatively large number of grid points from the outer region for supersonic flow computations. This enables us to save memory space since the memory is proportional to the total number of grid points. On the Cray computers, the eliminated grid points result in decreased computation time. However, the computation involved in non-reflecting type boundary conditions is relatively more involved than simple free-stream type boundary conditions. The overhead imposed by complex boundary conditions is relatively low on Crays compared to the CM.

For Mach = 1.5 and a 3° angle of attack, the solution clearly shows the asymmetry in the shock wave in the nose area, the region of over-expansion and recompression [Figure 2]. Notice that the use of non-reflecting boundary conditions allow the shock to pass through the outer boundary. A sequence of frames from the visualization of the computed flow fields reveals unsteady vortex shedding with vortices being formed and shed in the wake which raises questions about dislocation of events. Thus real-time visualization has opened up an area that needs further investigation. It is interesting to note that unsteady flow behavior has been reported in [Ref. 5] and [Ref. 6] which was not previously noticed in experiments. As the Mach number is increased, the recirculation region becomes entrapped in the base area. Again the real-time visualization may prove a key element in this type of study.

Two three-zone grids have been used in this study. The grid dimensions used for the first grid are (30x94x39), (249x65x39) and (120x46x39) respectively. The second three-zone grid has dimension size of (32x128x64), (256x128x64), and (128x64x64). The performance of the entire code for the M549 on the CM with 32K processors or 1048 nodes is about 1400 Mflops. This compares to performance of the same application on a Cray-2 at 90 Mflops using a single processor and 340 Mflops on using four processors. A Cray-2 was used because of the large memory required for the 3-D applications.

The entire M549 application, including various boundary conditions and inter-zone communications takes about 12 sec/iteration on a 32K processor machine. The detailed breakdown of the timing reveals that the code spends about 4 sec/iteration solving the kernel (interior grid points)

and 8 sec/iteration applying the boundary conditions including zonal interface data transfer (or working on the boundary points). As discussed previously, the work involved on the boundary can impose impressive overhead. So there can be a huge gap when comparing the performance of the kernel with the performance of the entire code. In our case inter-zone communication has turned out to be one of the most expensive operations on the CM because of the long-distance communication involved. Still, the performance of the M549 application on the CM is very encouraging and more importantly there is room for future applications that involve larger grids. It is not difficult to come up with applications involving single-block grid and simple boundary conditions that can be executed concurrently with the interior points and obtain much better performance than what we have obtained for the M549 application. We are focusing on the issues that we encountered while keeping the code flexible and practical enough for a variety of applications.

We have performed 3-D flow computations on the CM for the U. S. Army Missile Command (MICOM). A missile configuration associated with a sled test flow field has been considered [Ref. 10]. The missile configuration has been gridded using a single zone grid of $256 \times 64 \times 64$ [Figure 3]. The high Reynolds number computations have been performed for the Mach 1.9 case at a 0° angle of attack [Figure 4].

An application involving a little more complex body shape is shown in Figure 5. This is the full F-15 aircraft configuration. Because of the body shape, the multi-zone gridding of this configuration is more involved and three institutions have been involved (National Science Foundation - Engineering Research Center, McDonnell Douglas Corp. and the U. S. Air Force) [Ref. 11]. Once the grid has been obtained, the computations about the entire aircraft configuration have been carried out using the Zonal code on the CM. The surface pressure distribution for F-15 at Mach = 1.5 and a 0° angle of attack is shown in Figure 6. A more interesting area of research would be to make a video similar to the M549, capturing a series of flow frames involving an aircraft maneuvering such that roll or pitch are changing. The visualization capabilities discussed in the next section may prove an indispensable tool for the above mentioned study. A detailed discussion of the F-15 results will be documented in an appropriate future report.

5. VISUALIZATION

Visualization of flow data demands rendering large quantities of data which will require improvements to the traditional techniques. While machines like the CM have attached frame buffers, these machines are typically located at a remote site making it desirable to transfer the data to a local machine for visualization. The massive amount of data makes it difficult to transfer files, particularly to different architectures or workstations with modest disk capacity.

One approach is to transfer the data to a local machine with sufficient disk capacity and do the visualization almost entirely in software. This allows the visualization to be accomplished on machines without special graphics hardware. If the transfer and visualization steps are separated, the computation, visualization, and storage can happen concurrently, and each process can be performed on the most appropriate architecture.

For these reasons, the visualization of flow data from the Connection Machine has been broken into two steps: 1) Transferring the data [Ref. 13] and 2) Displaying and manipulating images. To allow flexibility in choosing the platforms for transfer and visualization, both steps are portable; relying on no special hardware or architecture specific software.

Each timestep of the Navier-Stokes calculation produces several scalar values at each grid point (density, X, Y, Z momentum and energy). Every tenth timestep is dumped to a file on the Data Vault. The filename, along with the size, shape, and normalizing values is written to a shared memory segment queue on the front-end processor. The Zonal code is then free to resume processing, resulting in a minimal impact on code performance.

A separate process, on another quadrant, reads the information from the shared memory queue, opens the Data Vault file, reads it into the CM, and normalizes the data. Then using RdT, a data transfer library, selected planes are transferred to a local processor. The data is transferred in eXternal Data Representation (XDR) format, developed by Sun Microsystems [Ref. 12] and stored on the local machine in host binary format. Once on the local machine, the data is visualized in an X11 window using ShAYD, a collection of hardware independent routines that is used to visualize curvilinear, multi-zone grids and their scalar values. Since all of the rendering is done in software,

massive grids can be handled, one small section at a time, to produce a final image. Each scalar value is mapped to a hue and a light source can be optionally applied. The hue is interpolated across each polygon and used, along with the light value, to produce either an value in the X11 color map or a 24 bit RGB value. The 24 bit output is used to produce videos of many timesteps.

6. SUMMARY

A parallel multi-zone, three dimensional, Navier-Stokes code has been developed on the Connection Machine CM for computing steady and unsteady flows. Results are presented which clearly demonstrate that real-world applications in one area can be handled efficiently on this massively parallel architecture. The applications also demonstrate the viscous flow computations at high Reynolds number. The graphics capabilities developed in conjunction with the parallel computation are very useful for visualizing the flow and debugging the computation.

The performance of the algorithm results more from simplicity and data management than from sophisticated numerics. We have highlighted issues that we encountered in addressing real-world applications on the CM machine. There can be a big gap in the performance of different applications depending on the type of boundary conditions and long-distance communications involved.

The basic theme of the research effort is simplicity. By putting simple modules together in an appropriate way and avoiding unnecessary complexity (the law of "diminishing returns"), we have produced a useful and efficient algorithm on the massively parallel architecture.

7. FIGURES

M549 Computational Grid

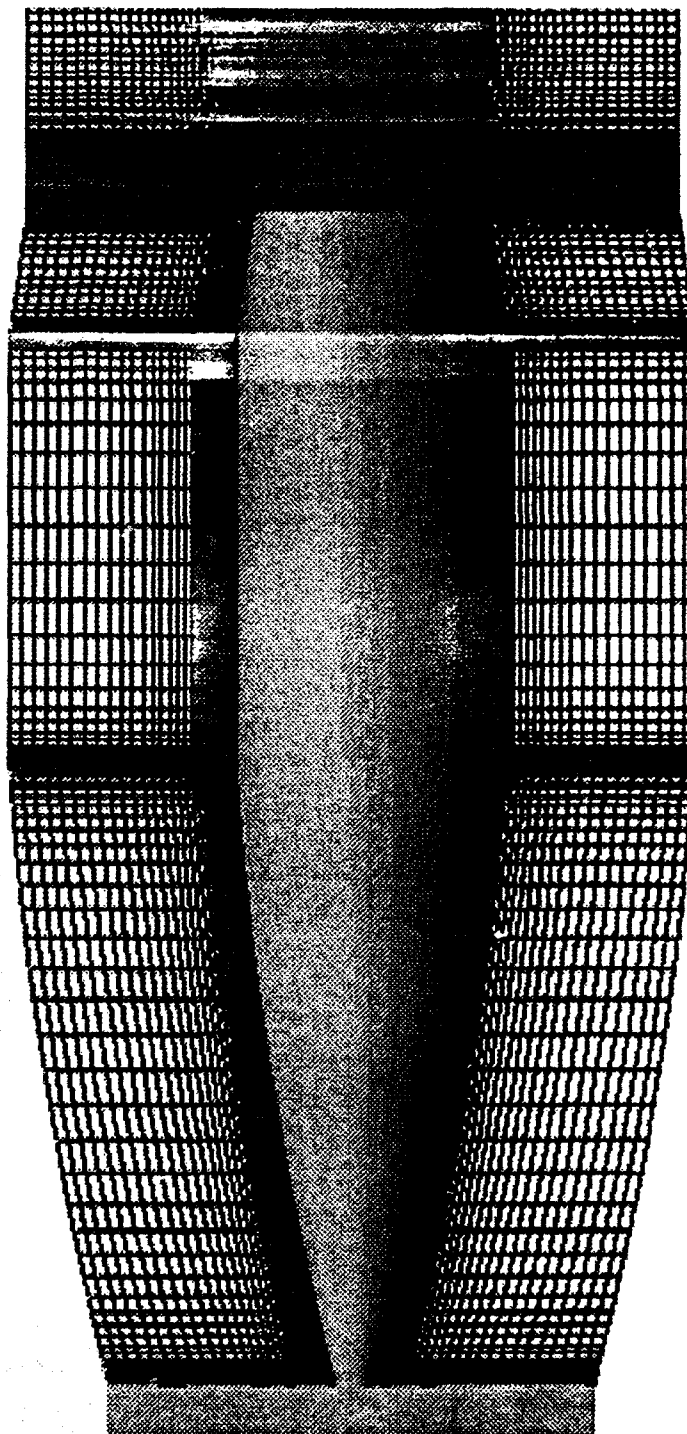


Figure 1

M1549 Density Contours

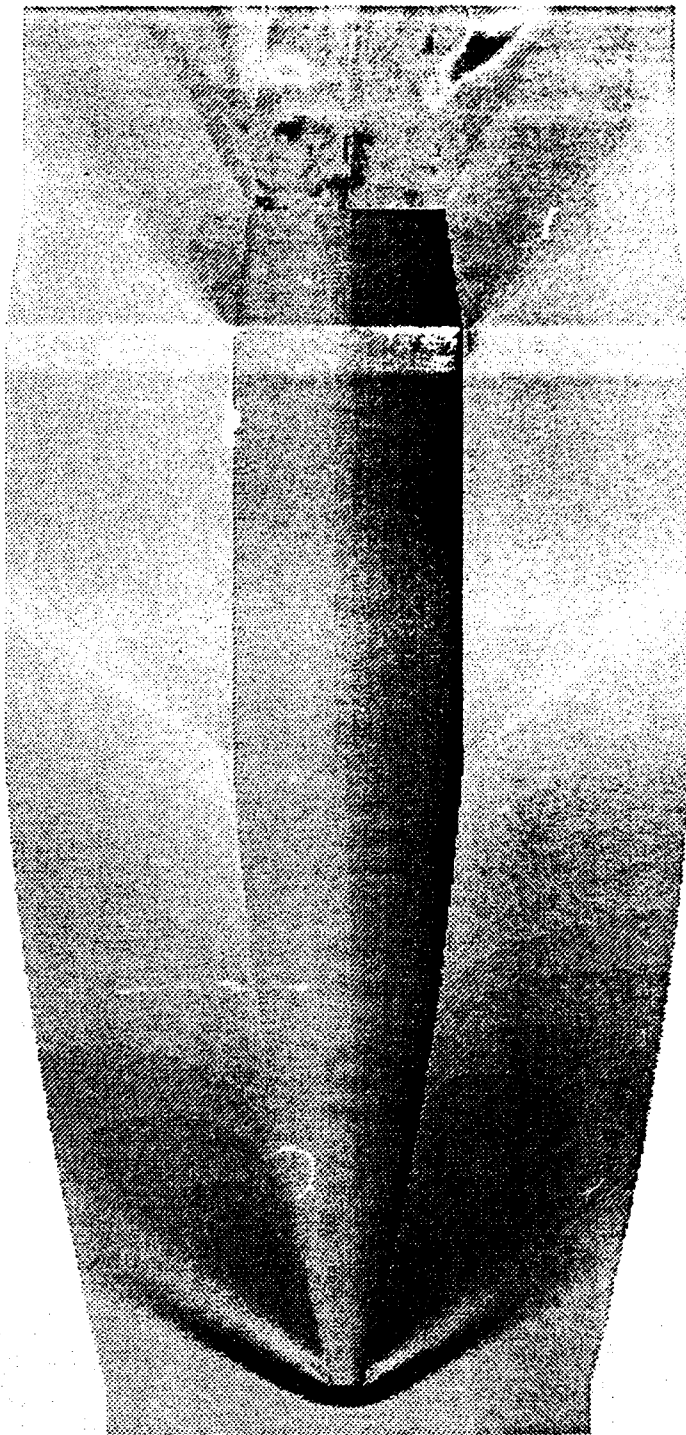
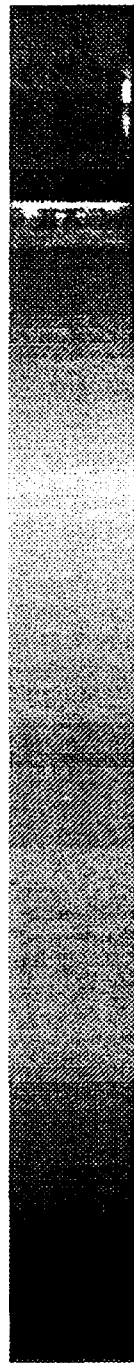


Figure 2



Figure 3

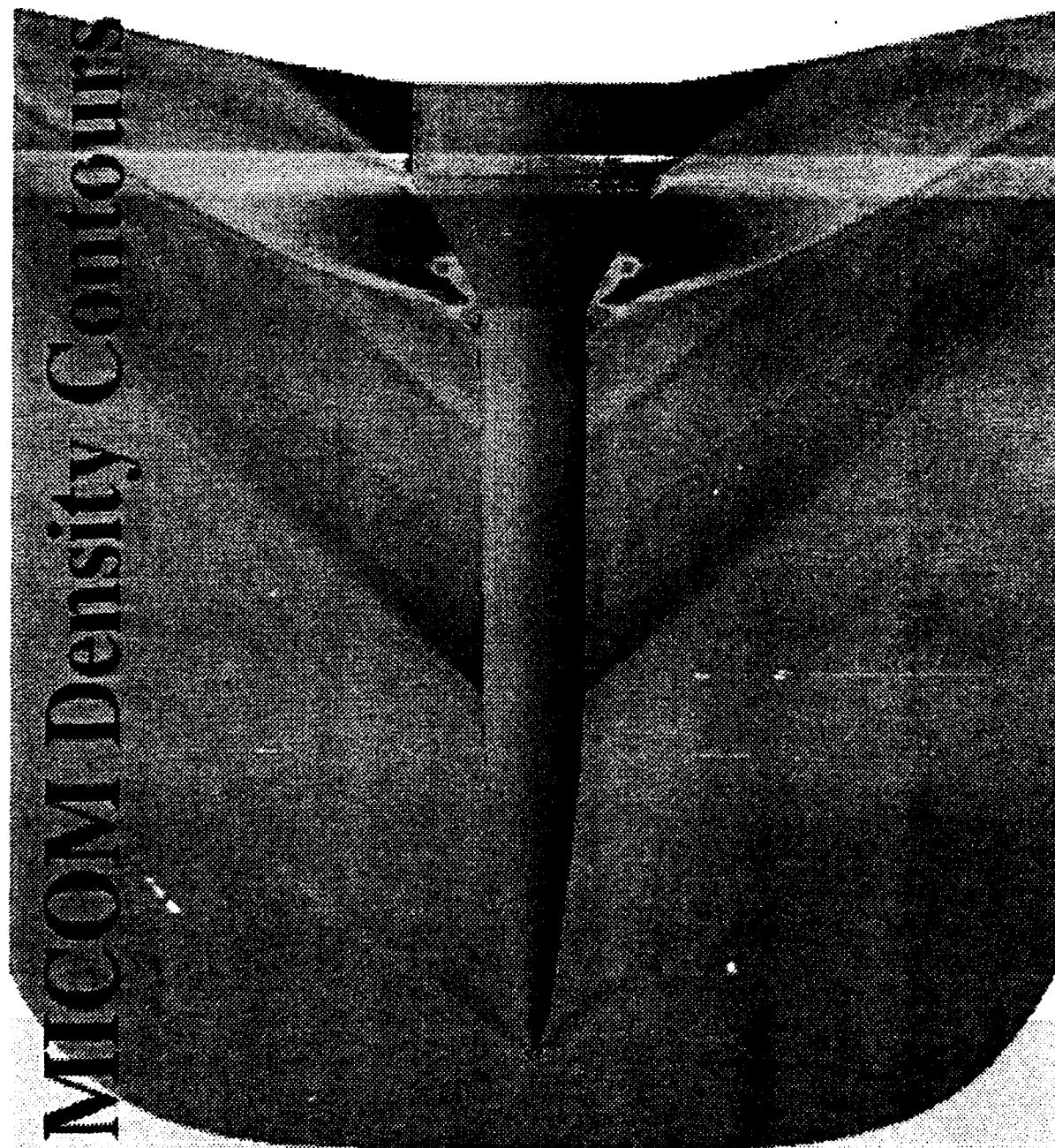
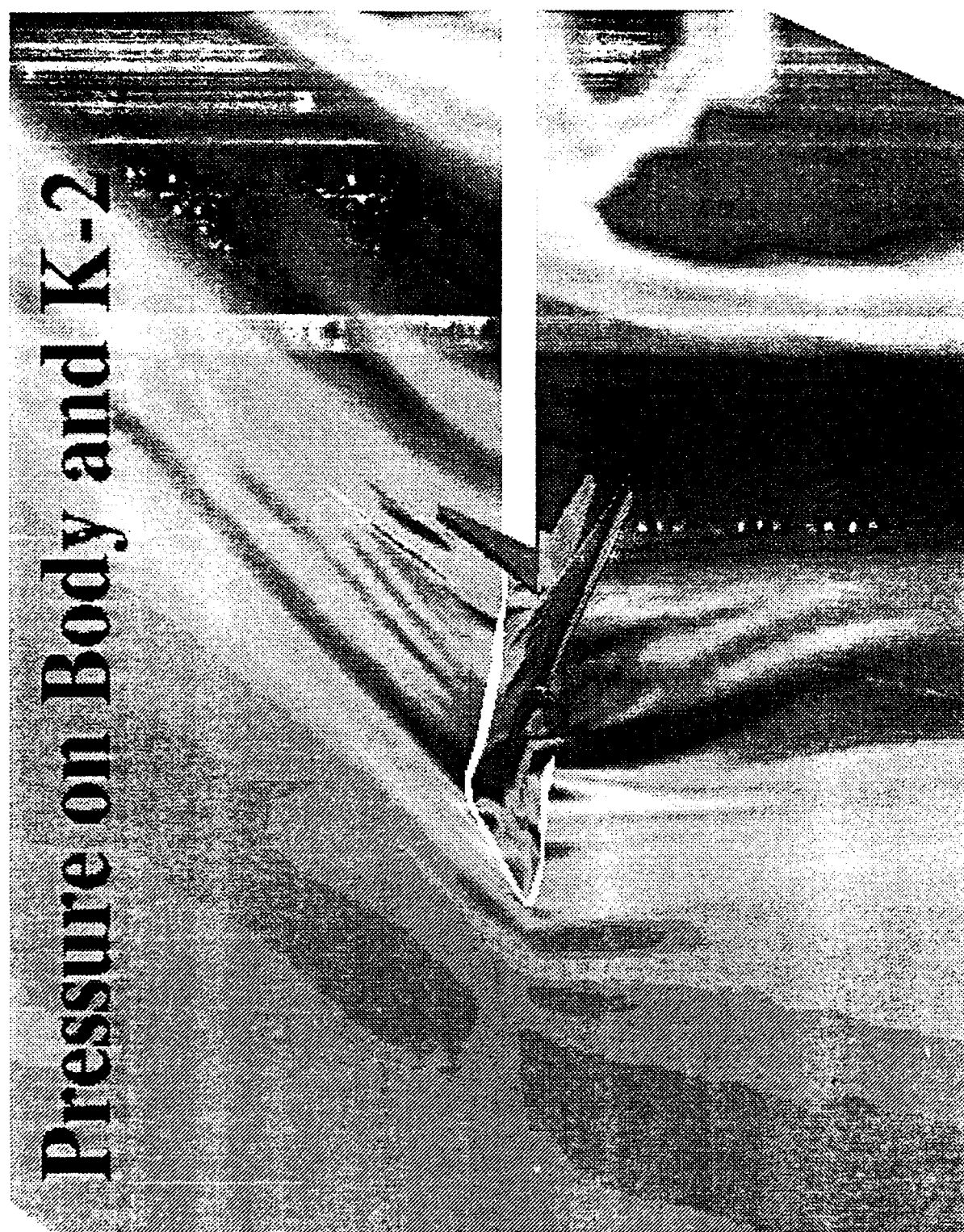


Figure 4

Pressure on F-15 Body



Figure 5



Pressure on Body and K-2

Figure 6

8. REFERENCES

- (1) Patel, N. R., Sturek, W. B. and Smith, G. A., "Parallel Computation of Supersonic Flows Using a Three Dimensional, Zonal, Navier-Stokes Code", BRL-TR-3044, November 1989.
- (2) Swanson, R. C. and Turkel, E., "Artificial Dissipation and Central Difference Schemes for the Euler and Navier-Stokes Equation", AIAA 8th Computational Fluid Dynamics Conference, Honolulu, Hawaii, June 1987.
- (3) Patel, N. R., Sturek, W. B. and Hiromoto, R. E., "A Parallel Compressible Flow Algorithm for Multiprocessors", Applications of Parallel Processing in Fluid Mechanics, Fluids Engineering Division, Vol. 47, ASME, June 1987.
- (4) Patel, N. R., Sturek, W. B. and Hiromoto, R. E., "A Parallel Numerical Simulation for Supersonic Flows Using Zonal Overlapped Grids on Common and Distributed Memory Multiprocessors", International Conference of Applications of Supercomputing in Engineering, Elsevier, Amsterdam, 1989, pp. 89-104.
- (5) Mikhail, Ameer, A., "Spike-Nosed Projectiles: Computations and Dual Slow Modes in Supersonic Flight", AIAA paper 89-1820, June 1989. (Also published in the Journal of Spacecraft and Rockets, vol. 28, No. 4, July-August, 1991, pp 418-424. Also BRL-TR-3140, August 1990.)
- (6) Mikhail, Ameer, A., "Spike-Nosed Projectiles With Vortex Rings: Steady and Non-Steady Flow Simulation", AIAA paper 91-3261, September 1991. (Also accepted for Journal of Spacecraft and Rockets, Also BRL-IMR-962, April 1991.)
- (7) Edge, H. L., Private Communication, BRL, June 1990.
- (8) Edge, H. L., "Computation of the Roll Moment Coefficient For a Projectile With Wrap-Around Fins", BRL-IMR-969, October 1991 (Also Proceedings of the Army Science Conference, June 1992).
- (9) Patel, N. R. and Edge, H. L., "Computations of Integrated Inlet-Combustor Flows for the National Aerospace Plane Engine", BRL-IMR-958, February 1991.
- (10) Soni, B. K., "Sled Test Flow Field", MICOM Technical Report no. 5-31915, 1992. (Also Private Communication.)
- (11) Soni, B. K., et al, Private Communication, March 1992.
- (12) "Networking on the Sun Workstation", Sun Microsystems, Mountain View, CA.
- (13) Clarke, J., "Remote Data Transfer (RdT): An Interprocess Data Transfer Method for Distributed Environments." BRL-TR-3339, May 1992.

**APPLICATION OF FINITE ELEMENT, GRID GENERATION, AND
SCIENTIFIC VISUALIZATION TECHNIQUES
TO 2-D AND 3-D SEEPAGE AND
GROUNDWATER MODELING**

Fred T. Tracy
Information Technology Laboratory
US Army Engineer Waterways Experiment Station
Vicksburg, MS

and

Camille A. Issa
Associate Professor of Engineering Mechanics
Department of Aerospace Engineering
Mississippi State University
Mississippi State, MS

ABSTRACT

The flow of groundwater in the vicinity of certain critical military arsenals is of increasing importance because of environmental impact. Further, there are many techniques used in military applications that can be used to solve problems in other disciplines (knowledge transfer). Therefore, this paper describes new advances in the computational modeling of groundwater and seepage using the finite element method (FEM) in conjunction with tools and techniques typically used by aerospace engineers. First, an extension of the technique of generating an orthogonal structured grid to automatically generate a flow net for two-dimensional problems is presented. Second, a complete implementation of a three-dimensional (3-D) seepage/groundwater model is described where (1) grid generation is accomplished using the Eglin Arbitrary Geometry Implicit Euler (EAGLE) program developed by Eglin Air Force Base and Mississippi State University, (2) the seepage and groundwater analysis for either confined or unconfined flow, homogeneous or inhomogeneous media, and isotropic or anisotropic soil is done with no restriction on the FE grid or requirement of an initial guess of the free surface for unconfined flow problems, and (3) scientific visualization is performed using the program Flow Analysis Solver Toolkit (FAST) developed by NASA Ames. Special emphasis is placed on the proper development of boundary conditions at exit faces and wells for unconfined flow or unsaturated flow problems, which are often done incorrectly. Finally, examples showing both generated flow nets for 2-D problems and results for 3-D applications are presented.

INTRODUCTION

The modeling of seepage under dams and groundwater flow in aquifers is of significant interest. This becomes even more important in our modern times with increasing interest in the flow of pollutants. The unsolved environmental issues regarding our hazardous and toxic waste problems must be resolved, and significant resources must be placed on this effort. Some military bases are contaminated with hazardous waste that has entered the groundwater domain. A groundwater model that takes into account contaminant flow is therefore critical. The state of the art has advanced in various ways over the years to achieve better and better solutions. However, of unusual occurrence is the application of the tools that engineers in one discipline have developed to problems of other disciplines. What is said is, "We don't do it that way." Because of the authors' diverse background, a unique feature of the work in this paper is that the tools developed by structural and aerospace engineers are applied to a problem typically addressed by others.

Difficulty of Problem

One profound aspect of seepage and groundwater flow is the problem of modeling flow through materials of significantly different characteristics (permeabilities, hydraulic conductivities, etc.). This is compounded when unconfined, unsaturated, or multiphase flow exists. One example of this is an earth dam with a relatively impervious clay core and a highly pervious drain installed around it. The rest of the dam is composed of moderately porous material. Figure 1 shows an example with the soil properties given in Table 1.

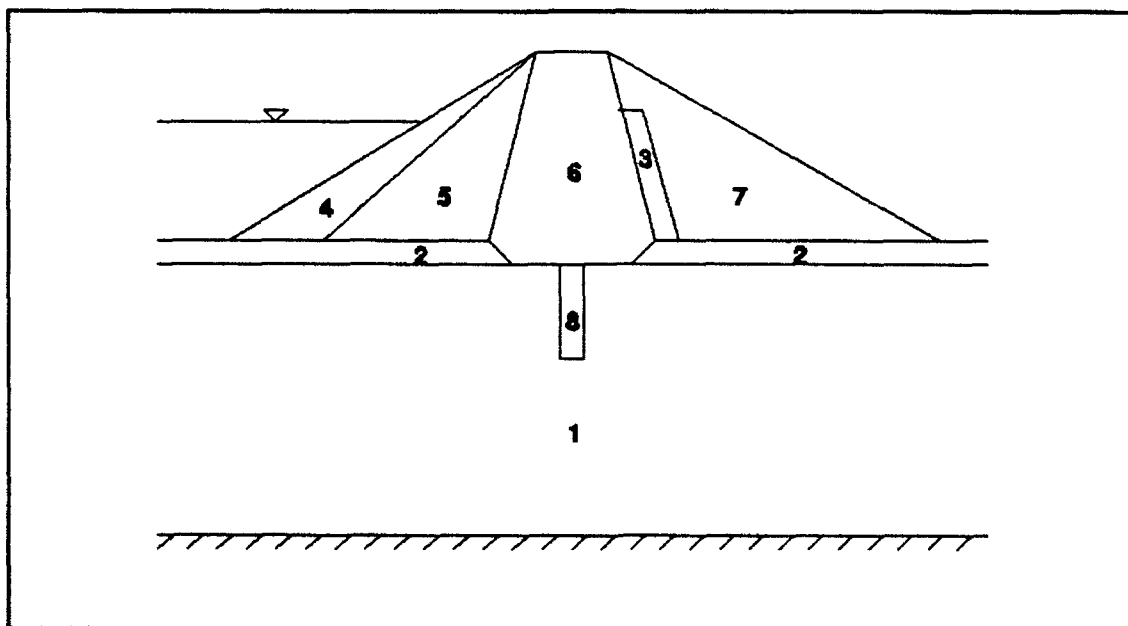


Figure 1. Zoned earth dam

PERMEABILITIES, FT/MIN				
Soil	Material	k_1	k_2	Angle, deg
1	Rock	$9.3(10^{-2})$	$1.7(10^{-2})$	140
2	Sand	$9.8(10^{-2})$	$2.0(10^{-2})$	0
3	Drain	$9.8(10^{-2})$	$2.0(10^{-2})$	0
4	Shell	$9.8(10^{-1})$	$9.8(10^{-1})$	0
5	Random	$9.8(10^{-3})$	$9.8(10^{-3})$	0
6	Core	$9.2(10^{-5})$	$2.0(10^{-5})$	0
7	Random	$9.8(10^{-3})$	$9.8(10^{-3})$	0
8	Grout	$9.8(10^{-3})$	$2.0(10^{-3})$	140

Table 1. Material properties

COMPUTER GENERATED FLOW NETS

The graphical construction of flow nets by hand to compute the quantity of flow, exit gradient, etc. is a standard engineering tool of soils engineers. However, these are extremely tedious to construct by hand because equipotential lines and flow lines must be drawn in such a way that curvilinear squares result. One significant aspect of this research effort is that numerical grid generation techniques of aerospace engineers used to generate an orthogonal grid (Thompson, Warsi, and Mastin 1985) can be extended to construct a flow net for various boundary conditions using the Cauchy-Riemann Equations (Crowder and McCuskey 1964). This section shows how the FEM has been successfully applied to generate flow nets with emphasis also given to differences in approach from previous work. The major advantage of the techniques described in this section is that they improve the quality of the resulting flow nets.

Governing Equations and Basic Approach

The total head or potential ϕ for a homogeneous, isotropic medium for 2-D, steady-state flow satisfies Laplace's equation as follows:

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} = 0 \quad (1)$$

The stream function ψ also satisfies Laplace's Equation,

$$\frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = 0 \quad (2)$$

Therefore, a complex potential Φ exists as follows:

$$\Phi = \phi + i\psi$$

Since ϕ and ψ are conjugate harmonic functions, the Cauchy-Riemann equations now hold.

$$\begin{aligned} \frac{\partial \phi}{\partial x} &= \frac{\partial \psi}{\partial y} \quad \text{or} \quad \phi_x = \psi_y \\ \frac{\partial \phi}{\partial y} &= -\frac{\partial \psi}{\partial x} \quad \text{or} \quad \phi_y = -\psi_x \end{aligned} \quad (3)$$

It should be noted here that the stream function is often defined as a velocity-type term. However, in this work it is a gradient-type term. That is, ψ is defined by

$$\psi = \int_c \phi_x dy - \phi_y dx$$

as compared to

$$\psi = \int_c u dy - v dx$$

where u is the x component of velocity, and v is the y component of velocity.

For a point on the curve C , a local tangent-normal (T-N) coordinate system can be established as shown in Figure 2. Equation 3 for this system then becomes

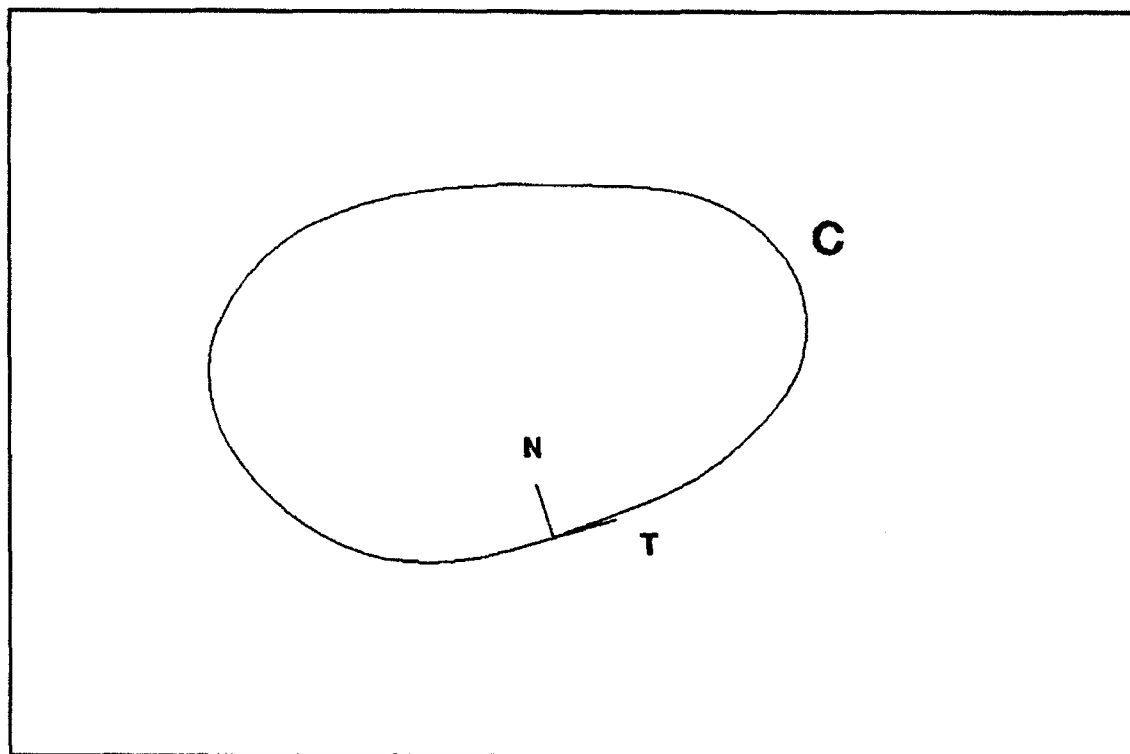


Figure 2. T-N coordinate system

$$\begin{aligned}\phi_T &= \psi_N \\ \phi_N &= -\psi_T\end{aligned}\tag{4}$$

A property of such functions is that constant lines of ϕ are orthogonal to constant lines of ψ . The flow net consists of $\phi = \text{constant}$ lines and $\psi = \text{constant}$ lines constructed in such a way that the resulting picture consists of curvilinear squares. The concept of automatically generating the flow net is fairly straightforward and involves the following three steps:

- a. *Compute total head.* Perform a normal FEM solution determining the total head h (same as potential ϕ) at each node and the quantity of flow, Q , passing through the system. Also compute the shape factor, f , from

$$Q = k(h_u - h_d) f\tag{5}$$

where h_u is the upstream head, h_d is the downstream head, k is the permeability, and f is the shape factor.

- b. *Compute stream function.* Determine the boundary conditions for the stream function using Equation 4 and perform a second FEM solution to obtain values of ψ at each node.
- c. *Contour results.* Contour the two sets of data to construct the flow net. The intervals for each are determined using the shape factor which, by definition, is

$$f = \frac{N_f}{N_e} \quad (6)$$

where N_e is the number of equipotential drops, and N_f is the number of flow paths.

Earlier work (Christian 1980a, 1980b, 1983; Aalto 1984; Christian 1987) determined the boundary values for the stream function solution (step 2) numerically, whereas in this work a more fundamental technique is used. Here, the Cauchy-Riemann equations are used to determine the correct boundary conditions.

Dupuit's Problem

The procedure will now be applied to Dupuit's problem which is unconfined flow in an earth dam with vertical sides (Figure 3). Line segment AB is impervious, line segments AF and BC have constant head specified, and line segment CD has the boundary condition

$$\phi = y \quad (7)$$

The position of the free surface (FD) must be determined from the FEM solution. Once determined, line segment FD becomes a flow line and is treated exactly like an impervious boundary. Also, the region above the free surface (triangular region FDE) is not used for the second solution. Rather, a new grid with the phreatic surface being a new boundary is used. We will now determine the ψ boundary conditions.

On the $\phi = H_1$ and $\phi = H_2$ boundaries

$$\phi_T = 0$$

since H_1 and H_2 are constants. But from Equation 4

$$\phi_N = 0$$

Applying Equation 4 to the above equation yields

$$\phi_N = -\psi_T = 0$$

or

$$\psi_T = 0 \quad (9)$$

Now

$$d\psi = \psi_T dT + \psi_N dN$$

Substituting Equation 9 into the above equation and noting that $dN = 0$ on the boundary gives

$$d\psi = 0$$

or

$$\psi = \text{constant}$$

The total amount of stream function can be shown to be

$$\psi_{total} = \frac{Q}{k} = f(h_u - h_d) \quad (10)$$

Therefore, apply the boundary conditions as follows:

$$\begin{aligned} \psi &= \psi_1 & \text{on } AB \\ \psi &= \psi_2 = \psi_1 + \psi_{total} & \text{on } FD \end{aligned} \quad (11)$$

where ψ_1 is an arbitrary constant set big enough so the FEM program has a confined flow condition when computing stream function.

On the boundary CD, Equation 7 is insufficient information to determine the new boundary conditions. Therefore, the normal component of discharge velocity v_N

is first computed for each node on the boundary CD with v_N at D being zero. It is assumed that points C and D are node points, and there are intermediate node points I and J (I is closest to C) as well. Then for each node,

$$v_N = -k\phi_N$$

$$\phi_N = -\psi_T$$

$$\psi_T = \frac{v_N}{k} \quad (12)$$

A piecewise linear assumption for v_N (and therefore ψ_T) yields

$$\psi_J = \frac{1}{2} (\psi_{TI} + \psi_{TJ}) (T_J - T_I) + \psi_I \quad (13)$$

Start I at point D and process consecutive nodes using Equation 13 until point C is done. The boundary conditions are now fully determined for the stream function calculation.

The above formulation is not restricted to Dupuit's Problem but can also be applied to a wide variety of quadrilateral-type earth dams. The only restriction is that the problem should have the five basic boundaries of impervious base, specified headwater and tailwater, free surface, and surface of seepage. Figure 4 shows the results for Dupuit's Problem where by Figure 3, $AB = AF = 100$ ft, and $BC = 20$ ft. Figure 5 shows the computer generated flow net for an earth dam.

SURFACE OF SEEPAGE BOUNDARY CONDITIONS

The free surface (FD in Figure 3) and the exit point (point D) must be determined by an iterative process, and the boundary conditions along CDE are often done incorrectly. What is typically done is a no-flow boundary condition is imposed along the entire boundary. This is incorrect because water must be allowed to exit along CD. So each iteration for the free surface must also include the following switches for all surface of seepage nodes:

- a. *Impervious node.* If the pressure is greater than zero, switch the node to a specified head node.
- b. *Specified head node.* If the flow Q is not flowing out, switch the node to an impervious node.

2-D SEEPAGE PACKAGE

A 2-D FEM seepage program with boundary conditions implemented correctly and with the flow net capability has been written and incorporated into a 2-D seepage package which contains (1) grid generation, (2) analysis, and (3) postprocessor capability. A PC version has been widely distributed and is available from the authors.

3-D MODEL

A 3-D seepage and groundwater model was created from the 2-D FEM program. Structured pieces of the 3-D grid for a problem are generated by EAGLE (Thompson 1987, Thompson and Gatlin 1988a, 1988b, 1988c) which has extensive algebraic grid generation capabilities, as well as state of the art elliptic smoothing capabilities. A program was then written to apply boundary conditions to individual blocks, combine blocks, and apply bandwidth minimization to preserve the element numbering. By decoupling the geometry into blocks, grid generation and boundary condition application is much less complicated. Also, the running time for solving the system of nonlinear equations can be significantly reduced by using a structured multi-block system. Further, the original structure can be preserved to output results in the Flow Analysis Solver Toolkit (FAST) (Bancroft, Kelaita, McCabe, Merritt, Plessel, Globus, and Semans 1991) format for scientific visualization.

Scientific Visualization Examples

The problem of two partially penetrating wells in a rectangular aquifer (Figure 6) bounded by a river and three impervious boundaries will first be given to show how groundwater can be visualized.

An analytic solution to a partially penetrating well (Figure 7) of thickness t_w , penetration b , well radius r_w , permeability k , and flow Q has been partially described (Muskat 1946), and the complete solution will now be given. First, define

$$\left\{ \begin{array}{l} \alpha = \frac{r}{2t_w} \\ \zeta = -\frac{z}{2t_w} \\ \beta = \frac{b}{2t_w} \end{array} \right. \quad (14)$$

where $(r, z) = (0, 0)$ is located at the top of the aquifer and at the center of the well. Next, define the function

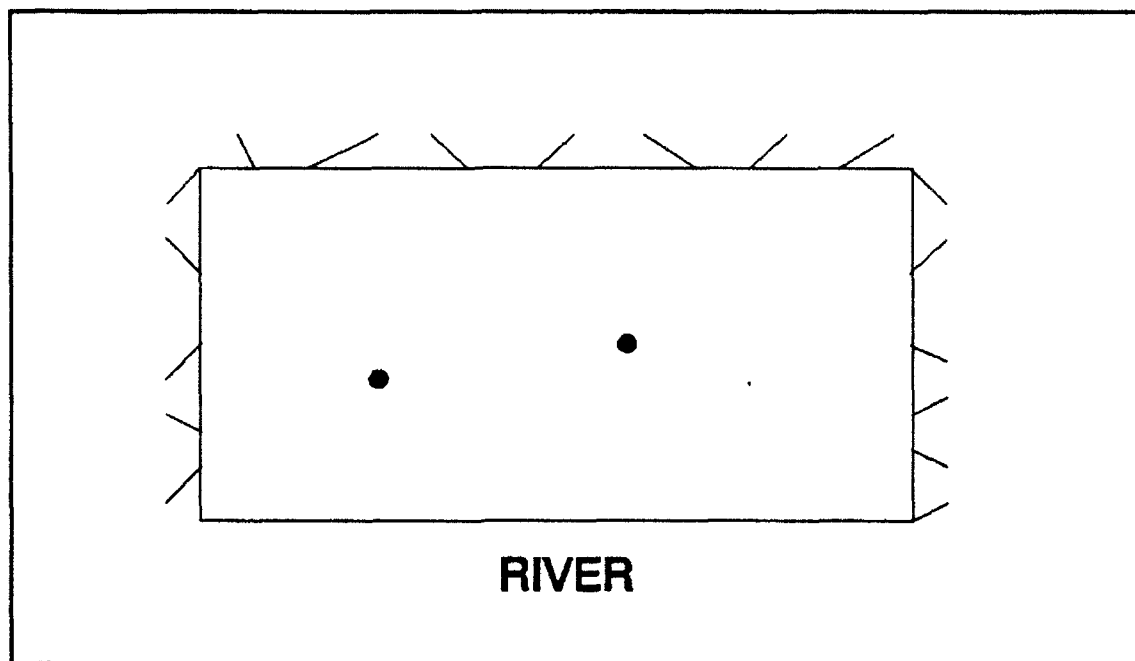


Figure 6. Two wells in an aquifer

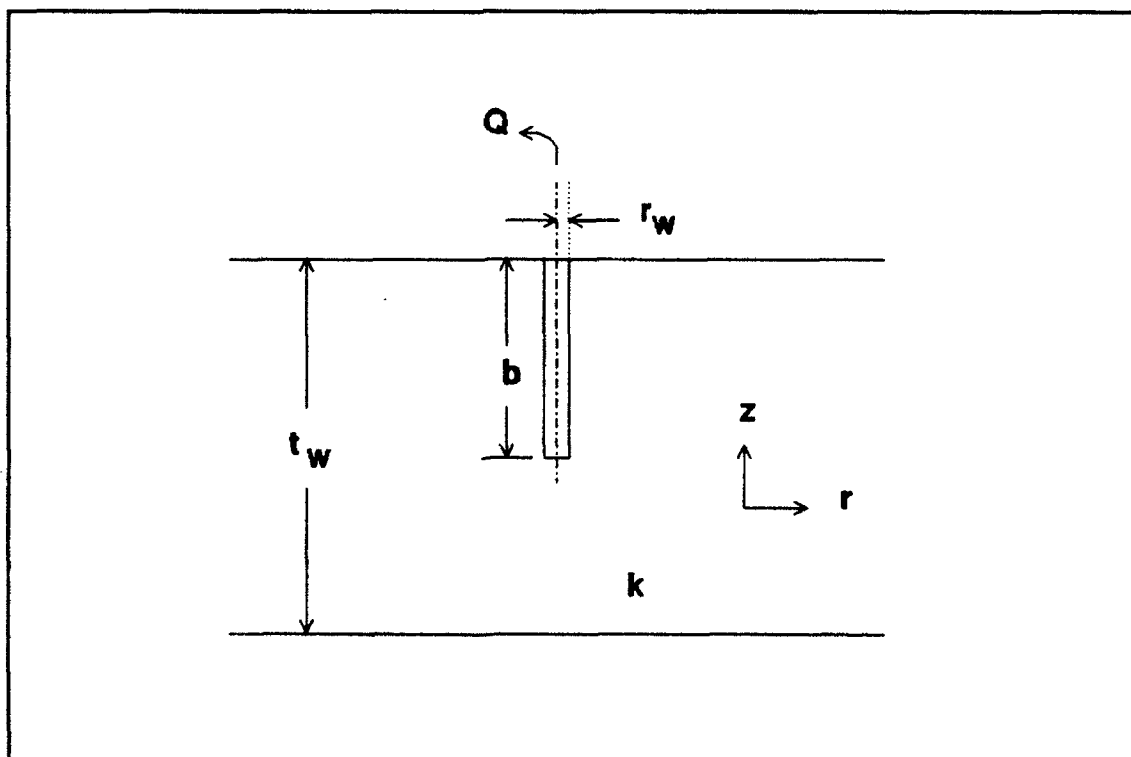


Figure 7. Partially penetrating well

$$Z(s, y) = \sum_{n=0}^{\infty} \frac{1}{(n+y)^s} \quad (15)$$

For small values of α ,

$$0 \leq \alpha \leq 0.5$$

the total head, ϕ , becomes

$$\begin{aligned} \phi = & q \left\{ -\log \frac{\Gamma(1+\zeta+\beta)\Gamma(1-\zeta+\beta)}{\Gamma(1-\zeta-\beta)\Gamma(1+\zeta-\beta)} \right. \\ & + \log \frac{\zeta+\beta+\sqrt{\alpha^2+(\zeta+\beta)^2}}{\zeta-\beta+\sqrt{\alpha^2+(\zeta-\beta)^2}} \\ & - \frac{\alpha^2}{4} [Z(2, 1-\zeta-\beta) - Z(2, 1-\zeta+\beta) \\ & + Z(2, 1+\zeta-\beta) - Z(2, 1+\zeta+\beta)] \\ & \left. + O(\alpha^4) \right\} \quad (16) \end{aligned}$$

q is a flux density given by

$$q = - \frac{Q}{4\pi kb} \quad (17)$$

For the remaining larger values of α

$$\begin{aligned} \phi = & 4q \left[\frac{1}{\pi} \sum_{n=1}^{\infty} \frac{i}{n} K_0(2n\pi\alpha) \cos(2n\pi\zeta) \sin(2n\pi\beta) \right. \\ & \left. + \beta \log \frac{2}{\alpha} \right] \quad (18) \end{aligned}$$

where $K_0(x)$ is the modified Bessel Function of the second kind. For relatively large values of x (Press, Flannery, Teukolsky, and Vetterling 1989),

$$K_0(x) \approx \frac{\pi}{\sqrt{2\pi x}} e^{-x}$$

Thus, an implementation of Equation 18 requires relatively few terms.

An infinite series of images, the first eight being shown in Figure 8, is now used to determine the solution. If a well is located at (x_0, y_0) and the top of the

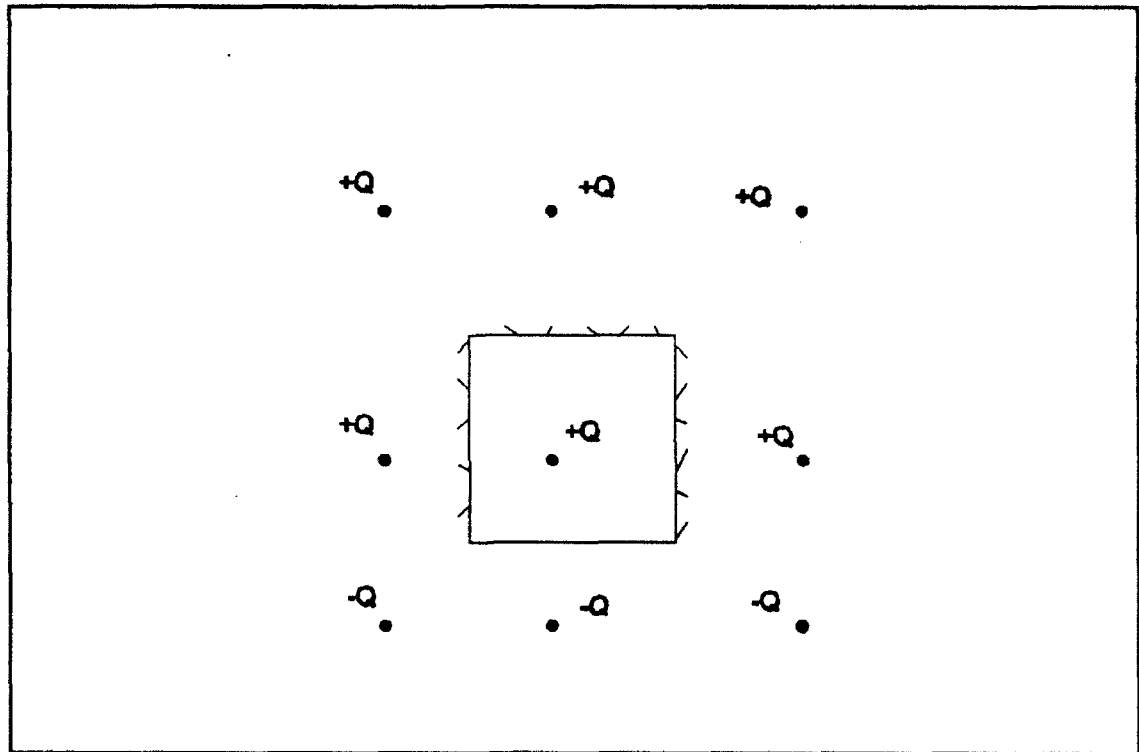


Figure 8. Image wells

aquifer is at z_t , then use

$$\alpha = \frac{1}{2h} \sqrt{(x - x_0)^2 + (y - y_0)^2}$$

$$\zeta = \frac{1}{2h} (z_t - z)$$

in Equations 16 and 18.

The above solution was done using 29 image wells on a large algebraic grid of 21,266 nodes and had a maximum percentage error of 5.3 percent. Actually, some of this error is attributed to the truncation in the number of image wells and some is due to numerical imperfections. It seemed plausible that this grid could be significantly reduced in size by applying the elliptic grid generation techniques of EAGLE to get close to the same result. A smaller elliptic grid of 13,266 nodes yielded a maximum percentage error of 5.7 percent which substantiates the hypothesis.

Figure 9 shows a K level shaded contour plot of potential for one of the wells with the highest potential being at the river (white) and the lowest potential being at the wells (black). Figure 10 shows several I level surfaces with shaded contour plots of potential. Figure 11 shows an isolevel plot, and Figure 12 shows a particle trace plot of flow from the river to the wells. One surprising result is that the left most flow line skips the smaller well and goes for the bigger one.

An extension of the first example is unconfined groundwater flow in a part of an aquifer (Figure 13) with a river, three partially penetrating wells, and a slurry trench (EF). Here, 16 subregions (13 going the full depth of the aquifer and 3 being below the partially penetrating wells) are used. Figure 14 shows a shaded contour plot of total head for this problem with the river painted dark for visualization purposes. It is interesting to see the effect of the impervious slurry trench on the head distribution.

CONCLUSION

The techniques and tools developed by aerospace engineers can be extremely helpful in groundwater flow problems. Also, visualization techniques are an essential and valuable tool for understanding and validating the results. An area of future research is the application of finite volume techniques used to solve aerospace problems being applied to multiphase flow of groundwater with contaminant transport.

ACKNOWLEDGMENTS

This work was sponsored by the WES In-House Laboratory Independent Research Program (ILIR) and the Engineering Division, Civil Works Directorate, Headquarters, US Army Corps of Engineers (HQUSACE), under the Computer-Aided Structural Engineering (CASE) Project. The Chief of Engineers, HQUSACE, has granted permission for the publication of this information.

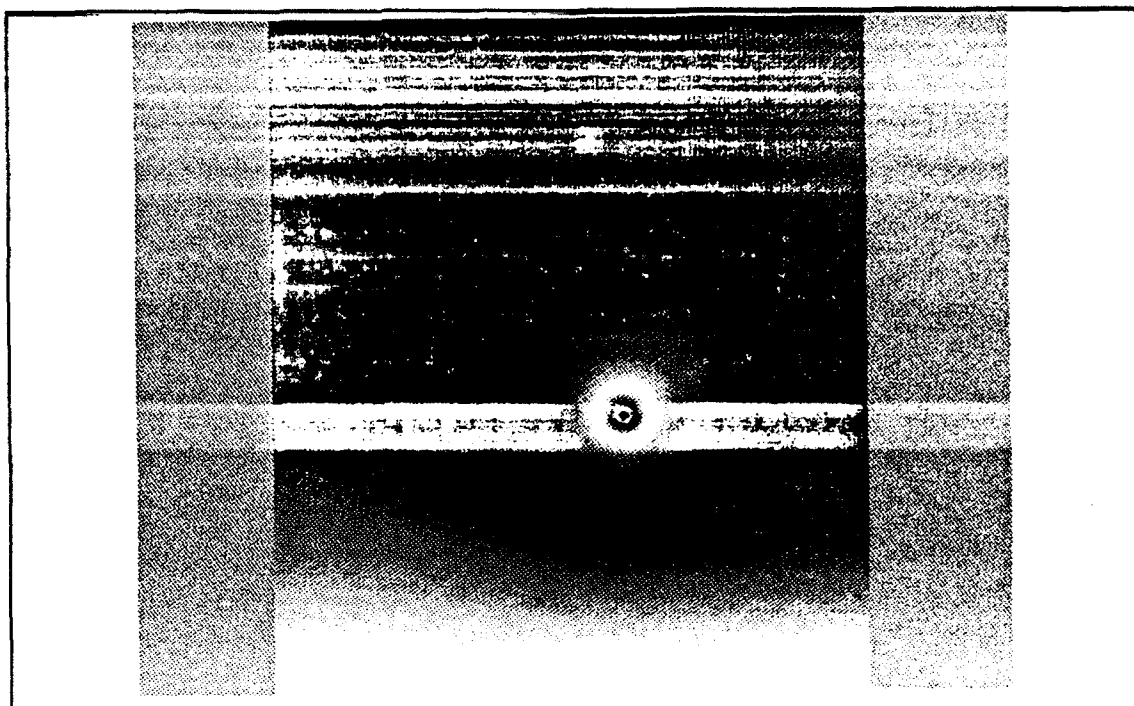


Figure 9. K level shaded contour plot

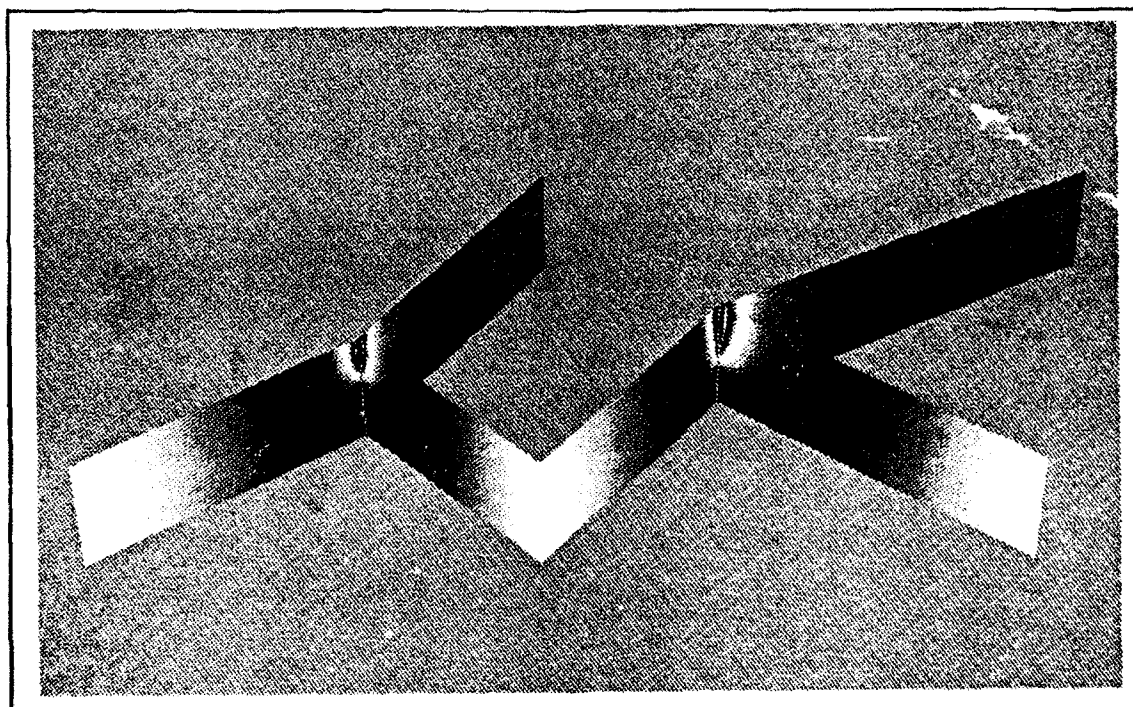


Figure 10. I level shaded contour plot

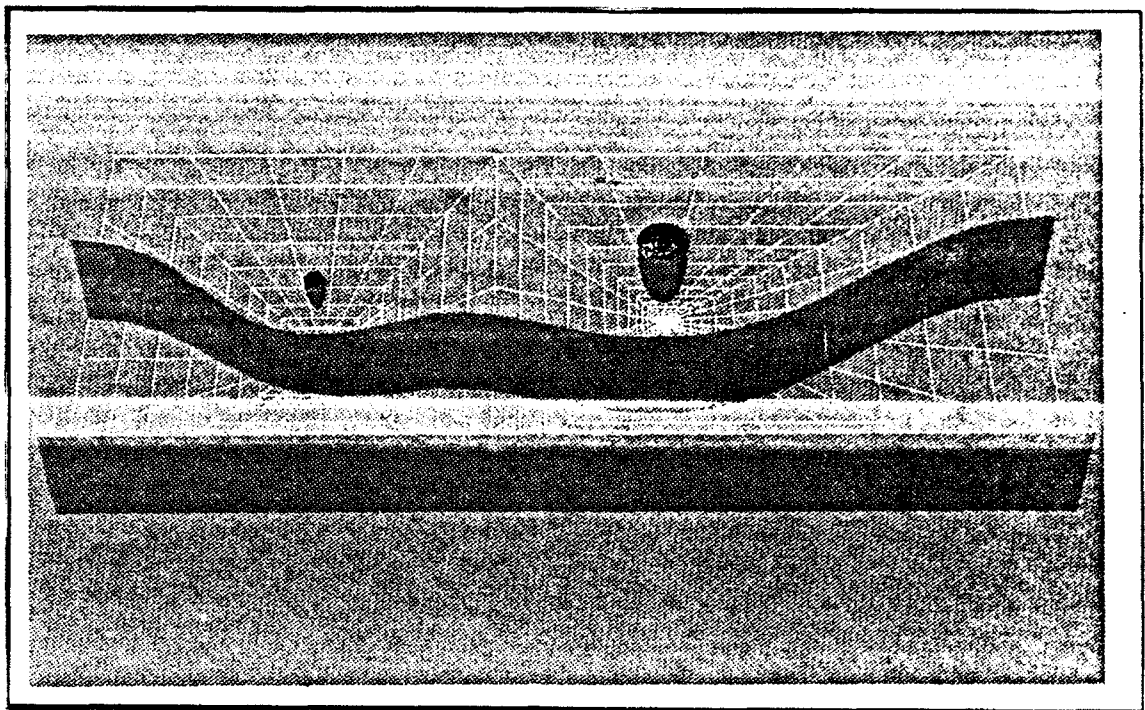


Figure 11. Isolevel plot

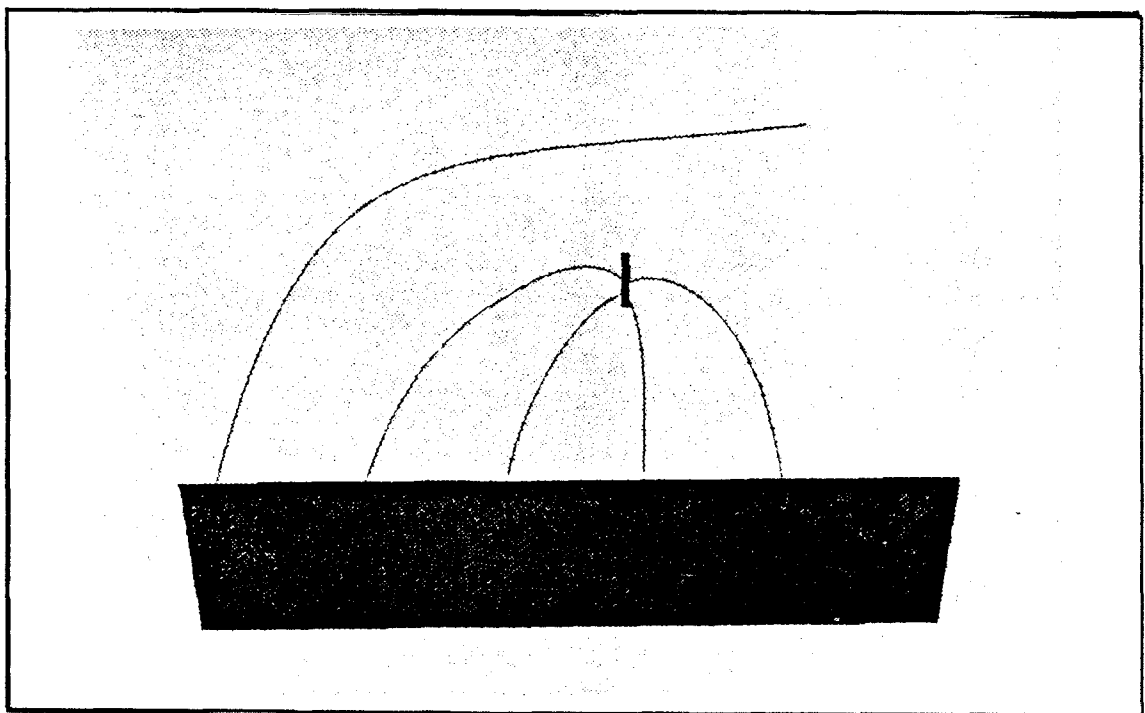


Figure 12. Particle trace plot

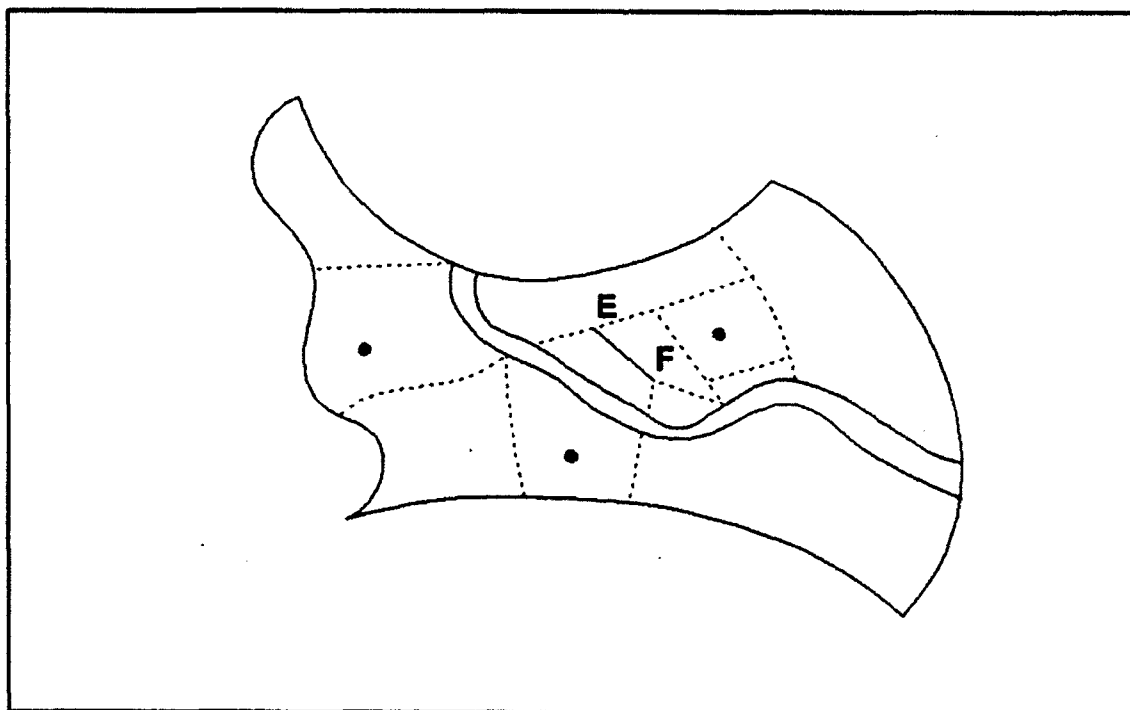


Figure 13. Part of an aquifer

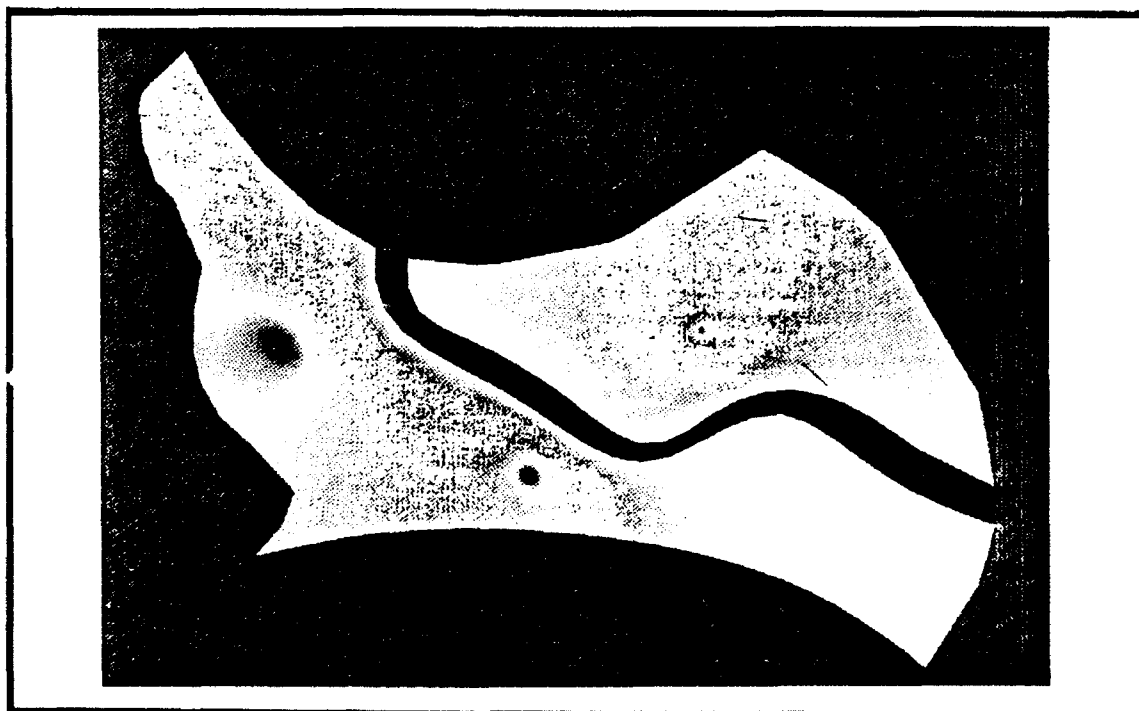


Figure 14. Shaded contour of total head

REFERENCES

- Aalto, J. (1984). Finite Element Seepage Flow Nets. *International Journal for Numerical and Analytical Methods in Geophysics*, 8, 297-303.
- Bancroft, Gordon, Kelaita, Paul, McCabe, Kevin, Merritt, Fergus, Plessel, Todd, Globus, Al, and Semans, John. (1991). Flow Analysis Software Toolkit (FAST). Sterling Federal Systems, Inc., NASA Ames Research Center, Moffett Field, CA.
- Christian, John T. (1980a). Flow Nets by the Finite Element Method. *Ground Water*, 18(2).
- _____. (1980b). Flow Nets form Finite Element Data. *International Journal for Numerical and Analytical Methods in Geomechanics*, 4(2).
- _____. (1983). Geotechnical Use of Finite Element Flow Analyses. Annual Convention of American Society Civil Engineers.
- _____. (1987). Establishing Boundary Condition for Stream Function Contouring of Finite Element Results. *Microsoftware for Engineers*, 3(2).
- Crowder and McCuskey. (1864). *Topics in Higher Analysis*, pp. 462-467. New York: The MacMillian Company.
- Muskat, M. (1946). *The Flow of Homogeneous Fluids through Porous Media*, pp. 263-268. Ann Arbor, MI: J. W. Edwards, Inc.
- Press, William H., Flannery, Brian P., Teukolsky, Saul A., and Vetterling, William T. (1989). *Numerical Recipes, The Art of Scientific Computing*, pp. 177. New York: Cambridge University Press.
- Thompson, J. F. (1987). A Composite Grid Generation Code for General 3-D Region. American Institute of Aeronautics and Astronautics 25th Aerospace Science Meeting. Reno, NV.
- _____. and Gatlin, B. (1988a). Program EAGLE User's Manual, Volume 1: Introduction and Grid Applications. USAF Armament Laboratory Technical Report AFATL-TR-88-117, Eglin AFB, FL.
- _____. and Gatlin, B. (1988b). Program EAGLE User's Manual, Volume 2: Surface Generation Code. USAF Armament Laboratory Technical Report AFATL-TR-88-117, Eglin AFB, FL.
- _____. and Gatlin, B. (1988c). Program EAGLE User's Manual, Volume 3: Grid Generation Code. USAF Armament Laboratory Technical Report AFATL-TR-88-117, Eglin AFB, FL.
- _____. Warsi, Z. U. A., and Mastin, C. Wayne. (1985). *Numerical Grid Generation, Foundation and Applications*, pp. 338-352. New York: Elsevier Science Publishing Co., Inc.

LIST OF CONFERENCE ATTENDEES

TENTH ANNUAL ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

16-19 JUNE 1992

NAME

Dr. Gerald R. Andersen
U.S. Army Research Office
Research Triangle Park, NC 27709

Joseph Arkin
Dept of Math Sciences
West Point, NY 10996

Harry J. Auvermann
US Army Atmospheric Sciences Lab
White Sands Missile Range, NM 88002

Marco Avellaneda
Courant Institute
251 Mercer St.
New York, New York 10012

Julia A. Baca
USAE Waterways Experimentation Station
3909 Halls Ferry Rd.
Vicksburg, MS 39180

Chandrajit L. Bajaj
Dept of Computer Science
Purdue University
West Lafayette, IN 47907

Romesh C. Batra
125 ME Annex
University of Missouri-Rolla
Rolla, Missouri

James T. Baylot
U.S. Army Waterways Experimentation Station
Vicksburg, MS 39180-6199

Abel J. Blanco
U.S. Army Atmospheric Lab
White Sands Missile Range, NM 88002

Dr. Jagdish Chandra
U.S. Army Research Office
Research Triangle Park, NC 27709

Peter Chen
Benet Laboratories
Armament Research, Development and Engineer Center

Jie Chen
Department of Mathematics
Chicago, Illinois 60680

Charles K. Chui
Department of Mathematics
Texas A & M University
College Station, TX 77843

Kenneth D. Clark
Army Research Office
Research Triangle Park, NC 27709

Herbert E. Cohen
AMSAA
Aberdeen Proving Ground, MD 21005

Monte W. Coleman
U.S. Army Ballistic Research Lab
Aberdeen Proving Ground, MD 21005

James W. Cooley
Department of Electrical Engineering
University of Rhode Island
Kingston, Rhode Island 02881

Dr. A. Brinton Cooper
U.S. Army Ballistic Research Lab
Aberdeen Proving Ground, MD 21005-5066

Joseph M. Coyle
Benet Laboratories
Watervliet Arsenal
Watervliet, NY 12189-4050

Terry Cronin
CECOM Signals Warfare Directorate
VHFS Warrenton, VA 22186

Benjamin Cummings
USA HEL
Aberdeen Proving Grounds, MD 21005-5001

John Dalbec
Dept of Mathematics
Cornell University
Ithaca, NY 14853

C de Boor
CMS University of Wisconsin
Madison, WI 53705

Dr. Aaron Das Gupta
U.S. Army Ballistic Research Lab
Aberdeen Proving Ground, MD 21005

Dr. Robert De Kinder
Army Atmospheric Sciences Lab
White Sands Missile Range, NM 88002-5501

John Dennis
Math Sciences Department
Rice University
Houston, Texas 77251-4094

Prof Arthur P. Dempster
Dept of Statistics
Harvard University
Cambridge, MA 02138

Peter J. Fazio
US Army Ballistic Research Lab
Aberdeen Proving Ground, MD 21005

Donald A. French
Dept of Mathematical Sciences
University of Cincinnati
Cincinnati, OH 45221-0025

Saul Gelfand
School of Electrical Engineering
Purdue University
West Lafayette, Indiana 47907

Frank Giordano
Department of Mathematical Sciences
United States Military Academy
West Point, NY 10996

James Glimm
University of Stony Brook
Stony Brook, NY 11794

John W. Grove
Dept of Applied Mathematics
University of Stony Brook
Stony Brook, NY 11794

Morton Gurtin
Department of Mathematics
Carnegie Mellon
Pittsburgh, PA 15213

Leonid Gurvits
Siemens Research Corporation
Plainsboro, NJ 08536

Bill Hager
Mathematics Department
University of Florida
Gainesville, Florida

Mark Hanisch
Center for Applied Math
Cornell University
Ithaca, NY 14853

Harumi Hattori
Department of Mathematics
West Virginia University
Morgantown, WV 26506

Prof L.F. Henderson
Dept of Applied Math and Stat
State University of New York
Stony Brook, NY 11794-3600

Dr. David W. Hislop
US Army Research Office
Research Triangle Park, NC 27709

Birkett T. Huber
Department of Mathematics
Cornell University
Ithaca, NY 14850

William Jackson
US Army TACOM
Warren, MI 48397-5000

Arthur Johnson
Army Materials Technology Lab
Watertown, MA 02172

Charles S. Jones
USAE Waterways Experimentation Station
Vicksburg, MS 39180

Aravind K. Joshi
Dept of Computer and Info Science
University of Penn
Philadelphia, PA 19104

Ken Kennedy
Computer Science Department
Rice University
Houston, TX 77251

Vipin Kumar
Computer Science Department
University of Minnesota
Minneapolis, MN 55455

G.S. Ladde
Dept of Mathematics
University of Texas
Arlington, TX 76019

Dimitris C. Lagoudas
Dept of Civil Engineering
RPI
Troy, NY 12180

Prof. Tong J. Lee
Dept of Electrical Engineering
Cornell University
Ithaca, NY 14853

Prof. Philippe Loustau
Dept of Mathematics
George Mason University
Fairfax, VA 22030

Edith H. Luchins
Department of Mathematics (VP)
United States Military Academy
West Point, NY 10996

Franklin Luk
Computer Science Department
RPI
Troy, NY 12180

C. Wayne Mastin
U.S. Army Waterways Experimentation Station
Vicksburg, MS 39180-6199

Kelley Mohrmann
Dept of Math Sciences
USMA
West Point, NY 10996

Joe Myers
Dept of Math Sciences
USMA
West Point, NY 10996

Donald H. Nelson
USAE Waterways Experimentation Station
Vicksburg, MS 39180

Anil Nerode
Mathematical Sciences Institute
Cornell University
Ithaca, NY 14853

Thomas J Pence
Dept of Materials and Mechanics
Michigan State University
East Lansing, MI 4824-122

Bradley Plohr
Dept of Applied Mathematics and Statistics
University of Stony Brook
Stony Brook, NY 11790

JVR Prasad
School of Aerospace Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332

Claudia J. Quigley
US Army Materials Technology Laboratory
Watertown, MA 02172-0001

C.R. Rao
Statistics Dept
Penn State University
University Park, PA 16802

J. Rastegar
Mechanical Engineering Department
SUNY at Stony Brook
Stony Brook, NY 11794

Fernando Reitich
Dept of Mathematics
Carnegie Mellon University
Pittsburgh, PA 15213-3890

John R. Rice
Computer Science Department
Purdue University
West Lafayette, Indiana 47907

Dr. Joseph M. Santiago
U.S. Army Ballistic Research Lab
Aberdeen Proving Ground, MD 21005

S. Sathananthan
Division of Science and Mathematics
Jarvis Christian College
Hawkins, TX 75765

Royce Soanes
Benet Weapons Lab
Watervliet Arsenal
Watervliet, NY 12189-4050

Douglas R. Sommerville
4348 Downs Square
Belcamp, MD 21017

Dr. H. Mete Soner
Dept of Mathematics
Carnegie Mellon University
Pittsburgh, PA 15213

Ram P. Srivastav
Applied Math and Statistics
SUNY-Stony Brook
Stony Brook, NY 11794-3600

Allan Steinhardt
Dept of Electrical Engineering
Cornell University
Ithaca, NY 14850

Michael Stillman
Mathematical Sciences Institute
Cornell University
Ithaca, NY 14853

Moss Sweedler
Mathematical Sciences Institute
Cornell University
Ithaca, NY 14853

David W. Sykora
U.S. Army Waterways Experimentation Station
Vicksburg, MS 39180

Dr. Boleslaw K. Szymanski
Dept of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

I. Tadjbakhsh
Department of Civil Engineering
RPI
Troy, NY 12180

F. Tangerman
Dept of Mathematics
SUNY at Stony Brook
Stony Brook, NY 11794

Thomas Ting
University of Illinois
Chicago, Illinois 60680

Fred T. Tracy
Waterways Experimentation Station
Vicksburg, MS 39180

Noel J. Walkington
Carnegie Mellon University
Pittsburgh, PA 15213

Feng Wang
Department of Applied Mathematics
University of Stony Brook
Stony Brook, NY 11794-3600

Richard A. Weiss
USACEWES
Vicksburg, MS 39180

J. R. Whiteman
BICOM, Inst of Comp Mathematics
Brunel University, UK

Dr. Daniel Willard
Deputy Under Secretary of the Army (Operations Research)
Washington, D.C. 20310-0102

Paul R. Woodward
Army High Performance Computing Research Center
University of Minnesota
Minneapolis, MN 55415

Dr. Julian J. Wu
U.S. Army Research Office
Research Triangle Park, NC 27709

R. Yalamanchili
US Army ARDEC
Flanders, NJ 07836

Stephen Yau
Department of Mathematics, Statistics, and Computer Science
University of Illinois at Chicago
Chicago, IL 60680

Dr. Bin Zhou
Statistics Dept
Penn State University
University Park, PA 16802

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified			1b. RESTRICTIVE MARKINGS		
2a. SECURITY CLASSIFICATION AUTHORITY			3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release: Distribution unlimited		
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE					
4. PERFORMING ORGANIZATION REPORT NUMBER(S) ARO Report 93-1			5. MONITORING ORGANIZATION REPORT NUMBER(S)		
6a. NAME OF PERFORMING ORGANIZATION Army Research Office		6b. OFFICE SYMBOL (If applicable) AMXRO-MCS	7a. NAME OF MONITORING ORGANIZATION		
6c. ADDRESS (City, State, and ZIP Code) P. O. Box 12211 Research Triangle Park, NC 27709-2211			7b. ADDRESS (City, State, and ZIP Code)		
8a. NAME OF FUNDING / SPONSORING ORGANIZATION AMSC		8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER		
8c. ADDRESS (City, State, and ZIP Code)			10. SOURCE OF FUNDING NUMBERS		
			PROGRAM ELEMENT NO	PROJECT NO.	TASK NO.
			WORK UNIT ACCESSION NO.		
11. TITLE (Include Security Classification) Transactions of the Tenth Army Conference on Applied Mathematics and Computing					
12. PERSONAL AUTHOR(S)					
13a. TYPE OF REPORT Technical Report		13b. TIME COVERED FROM 01-92 TO 02-93		14. DATE OF REPORT (Year, Month, Day) 1993 March	
15. PAGE COUNT 770					
16. SUPPLEMENTARY NOTATION					
17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)		
FIELD	GROUP	SUB-GROUP	Fluid and solid mechanics, mathematical physics and numerical methods, symbolic computation, control theory, and Stochastic techniques.		
19. ABSTRACT (Continue on reverse if necessary and identify by block number)					
(U) This is a technical report resulting from the Tenth Army Conference on Applied Mathematics and Computing. It contains most of the papers in the agenda of this meeting. These treat many Army applied mathematical problems.					
20. DISTRIBUTION / AVAILABILITY OF ABSTRACT <input checked="" type="checkbox"/> UNCLASSIFIED/UNLIMITED <input type="checkbox"/> SAME AS RPT <input type="checkbox"/> DTIC USERS			21. ABSTRACT SECURITY CLASSIFICATION		
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Francis G. Dressel			22b. TELEPHONE (Include Area Code) (919) 549-4319		22c. OFFICE SYMBOL AMXRO-MCS